Wiring the Microbial Web of Earth – New Approaches for Scalable Computational Prediction of Interpretable Microbial Ecosystem Structure

---

Dissertation

zur

Erlangung der naturwissenschaftlichen Doktorwürde

(Dr. sc. nat.)

vorgelegt der

Mathematisch-naturwissenschaftlichen Fakultät

der

Universität Zürich

von

Janko Tackmann

aus

Deutschland

Promotionskommission

Prof. Dr. Christian von Mering (Vorsitz)

Prof. Dr. Reinhard Furrer

Prof. Dr. Jeroen Raes

Zürich, 2019

Abstract

Microorganisms are the principal biotic driver of life on Earth. They shape virtually every aspect of the planet's biosphere, through both the maintenance of global biogeochemical cycles and via essential symbiotic relationships with multi-cellular organisms. Questions related to how individual microbial species form interacting communities (ecosystems)—with a drastic impact on their hosts and environments—are being studied with rapidly accelerating intensity by the Microbial Ecology field. Enabled through recent innovations in sequencing technologies, staggering amounts of knowledge are lately being generated, which already yielded fascinating insights: microorganisms have now been identified in even the most extreme environments, all across the globe, and intriguing connections between hosts and their microbiota are continuously being discovered, including for instance links to disease development and host behavior.

So far, insights have mainly been gained through comparatively straightforward, descriptive analyses of static microbial community snapshots. Such workflows employ for instance diversity-based comparisons of community profiles or the identification of community members that are strongly associated with a condition of interest (e.g. a disease or lifestyle factor). Less research has however focused on disentangling the underlying interaction structures, which dictate ecosystem dynamics and thus ultimately mold the observed community patterns. Elucidating these complex relationships would allow a system-level understanding of microbial communities and inform experiments aimed at mechanistic understanding.

Unfortunately, experimental validation of ecological interactions is currently impossible for all but the smallest communities and is furthermore restricted to microbes culturable under laboratory conditions, which constitute only a tiny fraction of the known diversity. Nonetheless, modern quantities of culture-independent microbial sequencing data offer a wealth of information to fuel computational prediction approaches and alleviate these shortcomings. In particular, such data can be mined for statistical co-occurrence or co-avoidance patterns—indicative of positive (mutualist, commensal) or negative (competitive, parasitic, predatory or amensal) ecological interactions—to enable the prediction of microbial interaction network models. Applying this approach to globally distributed sequencing data, covering

diverse habitats and conditions, would result in a model of the microbial web of Earth that could allow a first glimpse at global microbial interaction patterns.

Throughout the last decade, many methods have been proposed for the statistical prediction of ecological interactions. However, these approaches typically do not account for a variety of artifacts, including for instance shared ecological and environmental dependencies. Such artifacts are particularly widespread in global, heterogeneous data sets and thus severely hamper the ecological interpretation of networks inferred from such data. Moreover, current methods generally do not scale to modern (cross-study) sequencing data quantities, which seriously limits the comprehensiveness of predicted models.

In this thesis, I present a new approach to address these shortcomings: FlashWeave. The method uses a flexible Probabilistic Graphical Modeling (PGM) framework to infer direct associations. These predictions are depleted of indirect (i.e. spurious) associations and thus enable sparser and more interpretable ecosystem models. In contrast to the majority of current methods, FlashWeave furthermore scales to data sets with hundreds of thousands of samples and can explicitly integrate environmental and technical factors into model inference. We found that FlashWeave outperformed other approaches in recovering the structure of simulated microbial ecosystems and, additionally, surpassed them in detecting verified interactions within a real-world data set of marine sequencing samples. We furthermore used FlashWeave to predict the to date largest microbial interaction network of the human gastrointestinal tract, which revealed striking signals of potential biological relevance. These include for instance unusually pronounced phylogenetic assortativity, extensive interactions within the rare biosphere and novel mutualist hub species. Moreover, FlashWeave allowed us to infer a global cross-biome interaction network, based on more than half a million sequencing samples that cover highly diverse habitats. In-depth analysis of this network in future studies promises interesting ecological insights.

In the second part of this thesis, I present a parallel line of work that demonstrates how biomarker discovery can also strongly benefit from the removal of indirect associations. In this context, spurious associations may appear between microbes and non-microbial variables of interest (driven, for instance, by ecological microbe-microbe interactions) and can result in numerous redundant biomarkers, which negatively impact prediction quality and complicate biological interpretation. We found that FlashWeave, applied to the exemplary task of identifying microbes directly association to a selection of human body sites, generated highly parsimonious and interpretable biomarker sets. The resulting biomarkers furthermore yielded outstanding predictive performance on both pure and mixed body site microbiota.

The work presented in this thesis is a major step towards a better understanding of global microbial interaction trends, with potential applications for instance in

probiotics development, next-generation culturing efforts and ecosystem engineering. It furthermore highlights approaches for more parsimonious and interpretable biomarker discovery, which can be crucial for instance in clinical or forensic applications.

Zusammenfassung

Mikroorganismen sind die wichtigste biotische Voraussetzung für das Leben auf der Erde. Nahezu alle Aspekte der planetaren Biosphäre werden maßgeblich von Einzellern beeinflusst – sowohl durch die Aufrechterhaltung globaler biogeochemischer Kreisläufe als auch durch essenzielle symbiotische Beziehungen mit mehrzelligen Organismen. Wie sich individuelle mikrobielle Spezies zu interagierenden Gemeinschaften (Ökosystemen) zusammenschliessen – mit beträchtlicher Wirkung auf von diesen Gemeinschaften abhängige Wirte und Lebensräume – ist Gegenstand intensiver Forschung im Feld der mikrobiellen Ökologie. Jüngste Durchbrüche im Bereich der Sequenziertechnologie führten zu einem sprunghaften Anstieg neuer Erkenntnisse in diesem Gebiet, mit faszinierenden Einsichten: Mikroorganismen wurden mittlerweile weltweit in den extremsten Umgebungen identifiziert und verblüffende Verbindungen zwischen Wirten und ihren Mikrobiomen werden zunehmend aufgedeckt, beispielsweise im Zusammenhang mit Krankenheiten oder veränderten Verhaltensmustern der Wirte.

Allerdings wurden diese Erkenntnisse in erster Linie anhand vergleichsweise einfacher, deskriptiver Analysen von statischen Momentaufnahmen mikrobieller Gemeinschaften gewonnen. Solche Untersuchungen umfassen zum Beispiel diversitätsbasierte Vergleiche unterschiedlicher Mikrobiome oder die Identifikation von Spezies mit starker Assoziation zu wichtigen nicht-mikrobiellen Faktoren (z.B. mit Bezug auf Krankheiten oder den Lebensstil ihrer Wirte). In weitaus geringerem Maße hat sich die Forschung bisher hingegen mit zugrunde liegenden Interaktionsstrukturen befasst, welche die Dynamik der Ökosysteme diktieren und damit letztendlich auch die beobachteten Mikrobiomprofile formen. Ein tieferer Einblick in diese komplexen Beziehungen würde das Verständnis von mikrobiellen Gemeinschaften auf der Systemebene ermöglichen, um beispielsweise die Entwicklung zielgerichteter Experimente zum besseren Verständnis grundlegender Mechanismen zu erlauben.

Derzeit ist die genaue experimentelle Prüfung ökologischer Beziehungen zwischen Mikroorganismen aber auf kleine Systeme mit wenigen Spezies beschränkt, welche zusätzlich im Labor kultivierbar sein müssen und damit nur einen Bruchteil der bekannten Diversität umfassen. Die heute verfügbaren Massen an kultur-unabhängigen Sequenzdaten bilden allerdings eine ausgezeichnete Grundlage für computerbasierte Vorhersageverfahren, welche diese Probleme umgehen. Solche Methoden umfassen

insbesondere Verfahren zur Erkennung von statistischen Kookkurrenz- und Vermeidungsmustern, welche Indizien für positive (mutualistische oder kommensale) oder negative (kompetitive, parasitäre, räuberische oder amensale) ökologische Interaktionen liefern. Aufgrund dieser computergestützten Vorhersagen können anschliessend Modelle von mikrobiellen Interaktionsnetzwerken entwickelt werden. Anwendung dieses Ansatzes auf global verteilte Sequenzdaten, welche hoch diverse Habitate und Bedingungen abdecken, würde die Vorhersage eines mikrobiellen Netzwerkes der Erde ermöglichen und damit erste Einblicke in globale mikrobielle Interaktionsmuster gewähren.

Viele Methoden zur Vorhersage von ökologischen Interaktionen aus Sequenzdaten sind im vergangenen Jahrzehnt entwickelt worden. Diese sind diese im Allgemeinen jedoch anfällig für eine Vielzahl methodischer Artefakte, bedingt unter anderem durch geteilte ökologische und umweltbezogene Abhängigkeiten. Genau diese sind in globalen, heterogenen Sequenzdaten aber allgegenwärtig und erschweren dadurch die ökologische Interpretation vorhergesagter Netzwerkmodelle erheblich. Desweiteren sind derzeitige Methoden zumeist nicht in der Lage, moderne Massen an studienübergreifenden Sequenzdaten zu verarbeiten, wodurch der Umfang der vorhergesagten Modelle maßgeblich eingeschränkt wird.

In dieser Dissertation präsentiere ich einen neuen Ansatz, welcher Lösungsstrategien für diese Probleme bietet: FlashWeave. Die Methode verwendet ein flexibles, auf Probabilistischen Graphischen Modellen (PGMs) basierendes Softwaresystem, welches die Vorhersage von direkten Assoziationen erlaubt. Diese sind weitgehend frei von indirekten (störenden) Assoziationen, wodurch die Vorhersage von besser überschaubaren und interpretierbaren Ökosystemmodellen möglich wird. Anders als die meisten derzeitigen Methoden, ist FlashWeave in der Lage, umfangreiche Sequenzdatensätze mit hunderten von tausenden Proben zu verarbeiten und erlaubt zusätzlich die Integration von umweltbezogenen oder technologischen Faktoren in Modellvorhersagen. Unsere Ergebnisse zeigen, dass FlashWeave bekannte ökologische Interaktionsstrukturen aus simulierten Datensätzen genauer vorhersagen kann als alternative Ansätze und desweiteren verbesserte Resultate für verifizierte Interaktionen in echten marinen Sequenzproben liefert. Durch FlashWeave konnten wir das bisher größte mikrobielle Interaktionsnetzwerk für den menschlichen Magen-Darm-Trakt entwickeln, welches auffällige, biologisch interessante Signale offenbart. Diese umfassen unter anderem eine ungewöhnlich ausgeprägte phylogenetische Assortativität, eine ausgedehnte Interaktionslandschaft innerhalb der raren Biosphäre und neuartige mutualistische Hub-Spezies. FlashWeave erlaubte uns ausserdem die Vorhersage eines globalen, mikrobiomübergreifenden Interaktionsnetzwerkes, basierend auf mehr als eine halben million Sequenzproben aus den unterschiedlichsten Habitaten. Tiefere Analyse dieses Netzwerkes in zukünftigen Studien könnte interessante ökologische

Einblicke offenbaren.

Im zweiten Teil dieser Dissertation präsentiere ich einen weiteren Arbeitszweig, in welchem wir zeigen konnten, dass auch die Identifikation von Biomarkern entscheidend von der drastischen Reduktion indirekter Assoziationen profitieren kann. Solche indirekten Assoziationen können hier zum Beispiel zwischen mikrobiellen Spezies und wichtigen nicht-mikrobiellen Faktoren entstehen (als Beiprodukt ökologischer Interaktionen zwischen mikrobiellen Spezies) und zu einer Großzahl an redundanten Biomarkern führen. Letztere können negative Folgen für die Vorhersagegenauigkeit haben und biologische Schlussfolgerungen erheblich erschweren. Wir haben dieses Phänomen anhand des exemplarischen Problems der Identifikation von mikrobiellen Spezies mit direkter Assoziation zu spezifischen menschlichen Körperstellen untersucht. Unsere Ergebnisse zeigen, dass FlashWeave klar interpretierbare, parsimone Biomarkersets identifiziert, welche ausserdem herausragende Klassifizierungsqualität für Proben mit sowohl reinen als auch vermischten menschlichen Mikrobiomen erreichen.

Diese Arbeit stellt einen wichtigen Schritt hin zu einem besseren Verständnis globaler mikrobieller Interaktionstrends dar, mit potenziellen Anwendungen unter anderem in der Entwicklung von Probiotika, neuartigen Kulturmethoden und Ansätzen im Ecosystem Engineering. Desweiteren zeigt sie neue Möglichkeiten für die Identifikation von besser interpretierbaren, parsimoneren Biomarker auf, welche unter anderem in klinischen oder forensischen Anwendungen von entscheidender Bedeutung sein können.

Acknowledgements

There are numerous people who made my years in Zürich special in many ways. First and foremost, I owe my deepest gratitude to Christian and João who—both in their own ways—provided me with all the inspiration, supervision and technical knowledge I needed to become not only a better scientist but, importantly, a more informed and independent thinker.

Christian, thank you for giving me the freedom and trust to pursue this exciting line of research. I'm continuously amazed by your ability to be interested and supportive of every new idea, however unusual, and your capacity of opening up new lines of thought even on unfamiliar topics. Your guidance is what enabled this work to grow into what it is now.

João, without our long discussions, from big-picture topics to even the tiniest technical implementation details, I would probably still be stuck figuring out one or the other thing. Your critical and sharp perspectives have profoundly shaped all parts of this work, some of which likely would not exist without it. I hope that I will ever be only half the mentor to someone as you were to me.

I am furthermore grateful to my thesis committee, Reinhardt Furrer, Jeroen Raes and Shinichi Sunagawa, for their thoughtful feedback on various aspects of this thesis. Without their insight, the work presented here would be much more incomplete.

Credit is also due to all collaborators who I had the pleasure of working with, in particular Natasha Arora. Natasha was always ready to discuss the latest findings, provided invaluable input on a world that was so foreign to me at the time (forensics), and held up a good spirit, even in the face of mixed results. The fruits of her unceasing motivation to push the project forward are reflected in the biomarker discovery part of this thesis.

I'd furthermore like to thank all the von Mering's, both present and past, who made the time in- and outside the lab a blast. I will always keep our nail-biting rounds at the kickers table, prolonged discussions about (literally) everything, and all the wacky distractions, including drone races, mud runs and zombie hunts, in good memory. You were the best group I could have hoped for!

I also want to highlight the wonderful scientific community at the IMLS. The retreats across the most beautiful sites in Switzerland and the fun times at the TGIFs (as well

as the Loch Ness) were always welcome diversions from scientific routine. You made going through this much more fun than it would have otherwise been.

Aber ganz besonders möchte ich mich bei meiner Familie für all die Unterstützung über viele Jahre bedanken, sowohl während des Doktorats als auch davor. Auch wenn der Abstand es nicht immer leicht macht, habt ihr mir doch immer die Kraft gegeben, auch in schwierigen Zeiten weiter meinen Weg zu gehen. Nichts in dieser Arbeit wäre schlussendlich ohne euch gedacht oder geschrieben worden: sie nicht nur mein Verdienst, sondern auch der eure.

Contents

Abstract **i**

Zusammenfassung **iv**

Acknowledgements **vii**

List of Abbreviations **xi**

List of Figures **xiii**

1 Introduction **1**

1.1 Microbial ecosystems 1

1.1.1 Natural microbial communities 1

1.1.2 Fundamental ecosystem properties 15

1.1.3 Microbial systematics and the species problem 22

1.2 Methods for studying microbial ecology 29

1.2.1 Culture techniques 29

1.2.2 Sequencing technologies 35

1.2.3 Bioinformatic preprocessing and phylotypes 39

1.2.4 Databases for microbial ecology 45

1.2.5 Standard methods for microbiome analysis 47

1.3 Computational inference of microbial ecosystem structure 52

1.3.1 Current tools for microbial interaction prediction 52

1.3.2 Spurious associations and their sources 57

1.3.3 Probabilistic Graphical Models as a framework for ecological network inference 62

1.4 Research goals 68

1.4.1 General aims 68

1.4.2 Current challenges 69

1.4.3 Proposed solutions 70

2 Manuscripts **72**
2.1 Rapid inference of direct interactions in large-scale ecological networks from heterogeneous microbial sequencing data 72
2.2 Ecologically informed microbial biomarkers and accurate classification of mixed and unmixed samples in an extensive cross-study of human body sites 113

3 Discussion **146**
3.1 Inference of microbial interaction networks 146
3.1.1 Improved network quality 146
3.1.2 Scalability achievements 148
3.1.3 Biological insights 150
3.1.4 Limitations and outlook 152
3.2 Ecologically informed biomarker discovery 156
3.2.1 Conceptual advantages 156
3.2.2 Application to human body site microbiota 157
3.2.3 Limitations and outlook 158
3.3 Concluding remarks 160

Appendices **161**

A MAPseq: highly efficient k-mer search with confidence estimates, for rRNA sequence analysis **161**

B Curriculum Vitae **176**

Bibliography **178**

List of Abbreviations

AIC:	akaike information criterion
ALR:	additive log-ratio transform
ANI:	average nucleotide identity
ARIMA:	autoregressive integrated moving average
BIC:	bayesian information criterion
BMI:	body mass index
BN:	bayesian network
CAMI:	critical assessment of metagenome interpretation
CLASI-FISH:	combinatorial labeling and spectral imaging-FISH
CLR:	centered log-ratio transform
CPR:	candidate phyla radiation
DBN:	dynamic BN
EMP:	earth microbiome project
ENA:	european nucleotide archive
ESV/ASV:	exact/amplicon sequencing variation
FBA:	flux balance analysis
FCI:	fast causal inference
FCM:	functional causal model
FDR:	false discovery rate
FISH:	fluorescence *in situ* hybridization
FPGA:	field-programmable gate array
gANI:	genomic ANI
GES:	greedy equivalence search
GFP:	green fluorescent protein
GLL:	generalized local learning
gLV:	generalized lotka-volterra
GPU:	graphics processing unit
HMP:	human microbiome project
ITS:	internal transcribed spacer
LGL:	local-to-global learning

LGT:	lateral gene transfer
MAG:	metagenome-assembled genome
MALDI-TOF:	matrix-assisted laser desorption with time-of-flight
MAP:	microbe atlas project
MAR-FISH:	microautoradiography-FISH
MB:	markov blanket
MCMC:	markov chain monte carlo
MDA:	multiple displacement amplification
MDL:	minimum description length
MGWAS:	metagenome-wide association study
MLST:	multilocus sequence typing
MRF:	markov random field
NanoSIMS:	nanoscale secondary ion mass spectrometry
NMDS:	non-metric multidimensional scaling
OTU:	operational taxonomic unit
PCoA:	principal coordinates analysis
PCR:	polymerase chain reaction
(P)DAG:	(partially) directed acyclic graph
PERMANOVA:	permutational analysis of variance
PGM:	probabilistic graphical model
RFE:	recursive feature elimination
SCFA:	short-chain fatty acid
SEM:	structural equation model
SIP:	stable isotope probing
SNV:	single nucleotide variation
SRA:	sequence read archive
SVM:	support vector machine
TINA:	taxa interaction-adjusted index
TPU:	tensor processing unit
WGS:	whole genome (shotgun) sequencing

List of Figures

1.1 The role of microorganisms in global biogeochemical cycles and as primary producers . . . 3
1.2 Important symbioses between microbial symbionts and animal or plant hosts . . . 5
1.3 The diverse roles of microbial endosymbionts in animal hosts . . . 8
1.4 Ecological interactions between microbes . . . 10
1.5 Ecological networks of plants and animals . . . 16
1.6 A selection of general properties found in complex networks . . . 17
1.7 Challenges and concepts for species delineation in microbiology . . . 27
1.8 Examples of culture and monitoring techniques . . . 31
1.9 Sequencing approaches for microbial community analysis . . . 37
1.10 Bioinformatic workflows for microbial community analysis . . . 41
1.11 Standard types of analyses for microbial abundance data . . . 49
1.12 Classes of methods for the prediction of microbial interaction networks 54
1.13 Sources of spurious associations in co-occurrence networks . . . 58
1.14 Bayesian networks and the GLL-LGL framework . . . 66
1.15 Proposed methods for inferring interpretable ecological networks and biomarkers from heterogeneous data sets . . . 70

Chapter 1
Introduction

1.1 Microbial ecosystems

1.1.1 Natural microbial communities

The role of the infinitely small in nature is infinitely great.

—Louis Pasteur

Earth is, as far as we currently know, the only planet to have spawned the most peculiar phenomenon in the known universe: life. It's atmosphere, surface and deep layers are teeming with organisms of astounding variety, spanning a multitude of scales. From the largest known life forms, the fungus *Armillaria ostoyae* (9 km^2 in area, Schmitt and Tatum (2008)) and the superorganism Pando (a colony of *Populus tremuloides*, 0.5 km^2 in area, 6600 t in weight (Mitton and Grant, 1996)), multicellular life reaches its current lower boundary in the tiny *Myxobolus shekeli* (about 8 µm in length, Kaur and Singh (2011)). From there, characterized diversity continues further into the unicellular domain, where species range from meters (*Caulerpa taxifolia*, with stolons reaching up to several meters in length (Meinesz et al., 1995)) all the way down to a few hundred nanometers (*Mycoplasma genitalium*, 200-300 nm in diameter (Moore, 1999)). Life as we know it thus spans a staggering scale of ten orders of magnitude.

Despite their size, unicellular organisms (microorganisms or microbes in short) are estimated to make up a major fraction of biomass on Earth (77 Gt C), vastly surpassing that of animals (2 Gt C) and being only second to plants (450 Gt C) (Bar-On et al., 2018). An extraordinary example is the alphaproteobacterial clade SAR11, which, at an estimated population size of 2.4×10^{28} cells (up to 50% of all cells in marine surface waters), is considered the most successful organism on Earth (Morris et al., 2002). Indeed, the number of extant microbial species is projected to be in the millions (Schloss et al., 2016) or possibly even in the trillions (Locey and Lennon, 2016), making them (potentially by far) the most diverse class of life forms on Earth. Microbes are found across all niches of our planet, including even exotic habitats—previously

thought to be devoid of life—such as deep-sea sediments (Zobell and Morita, 1959), hot acidic springs (Brierley and Brierley, 1973), the arctic permafrost (Wilhelm et al., 2012), human-made constructions in space (Vaishampayan et al., 2012), the deep terrestrial subsurface (Szewzyk et al., 1994) and many other extreme environments (Diels and Mergeay, 1990; Mormile et al., 2009; Rampelotto, 2013; Tirumalai et al., 2018). Remarkably, extremophiles are not only tolerant to multiple types of extreme environmental stresses (polyextremophiles), but indeed require them for proper growth (Rampelotto, 2013).

Another intrinsic trait of many microbial species is an unparalleled capacity for dispersal. For instance, microbial cultures of strict anaerobes can be obtained from aerobic samples (Bianchi et al., 1992) and thermophilic organisms can be grown from low-temperature marine samples (Isaksen et al., 1994), which shows the considerable spread of microbial cells and spores across diverse environments[1]. The high population size, variety of habitats, resilience to environmental stressors and potential of dispersal thus make microbes the most successful class of organisms on the planet.

This dominance in Earth's biosphere is reflected by the importance of microorganisms for other life forms: every organism is in some way dependent on them (see Figure 1.1). This becomes most evident through the planet's globally operating biogeochemical cycles (Williams, 1997), which provide storage, conversion and flow of the elements crucial for nutrition, homeostasis and maintenance of environmental conditions for Earth's species. Importantly, biogeochemical cycles are driven by the exceptional metabolic diversity found in microbial life: the carbon, nitrogen, phosphorus, sulfur and oxygen cycles all have major microbial steps contributing to their flow and balance (Falkowski et al., 2008; Madigan et al., 2014). Indeed, the loss of essential microbial drivers would cause these cycles to collapse, leading to the destruction of whole ecosystems (Doney et al., 2012; Treseder et al., 2012).

Not only were photosynthetic microbes the principal driver of the oxygenation of Earth (early cyanobacteria initiated the "Great Oxygenation Event" ca. 2090 and 2450 Ma ago, Farquhar et al. (2000)), they remain important O_2 contributors, responsible for approximately half of all global atmospheric oxygen production (Field et al., 1998; Walker, 1980). Similarly, bioavailable nitrogen is generated exclusively by microorganisms through the fixation of atmospheric nitrogen and conversion to NH_4^+ (Dixon and Kahn, 2004), which can subsequently be used by other organisms to construct complex nitrogen-containing organic compounds, such as nucleic acids and amino acids (see Figure 1.1 B).

Microbes are among the most important primary producers, i.e. initial converters of inorganic compounds into biologically available and energy-rich molecules at the lowest trophic level. An example of this is *Prochlorococcus*, a genus of marine cyanobacteria

[1] Recent human activity likely further promoted this intrinsic tendency (Zhu et al., 2017).

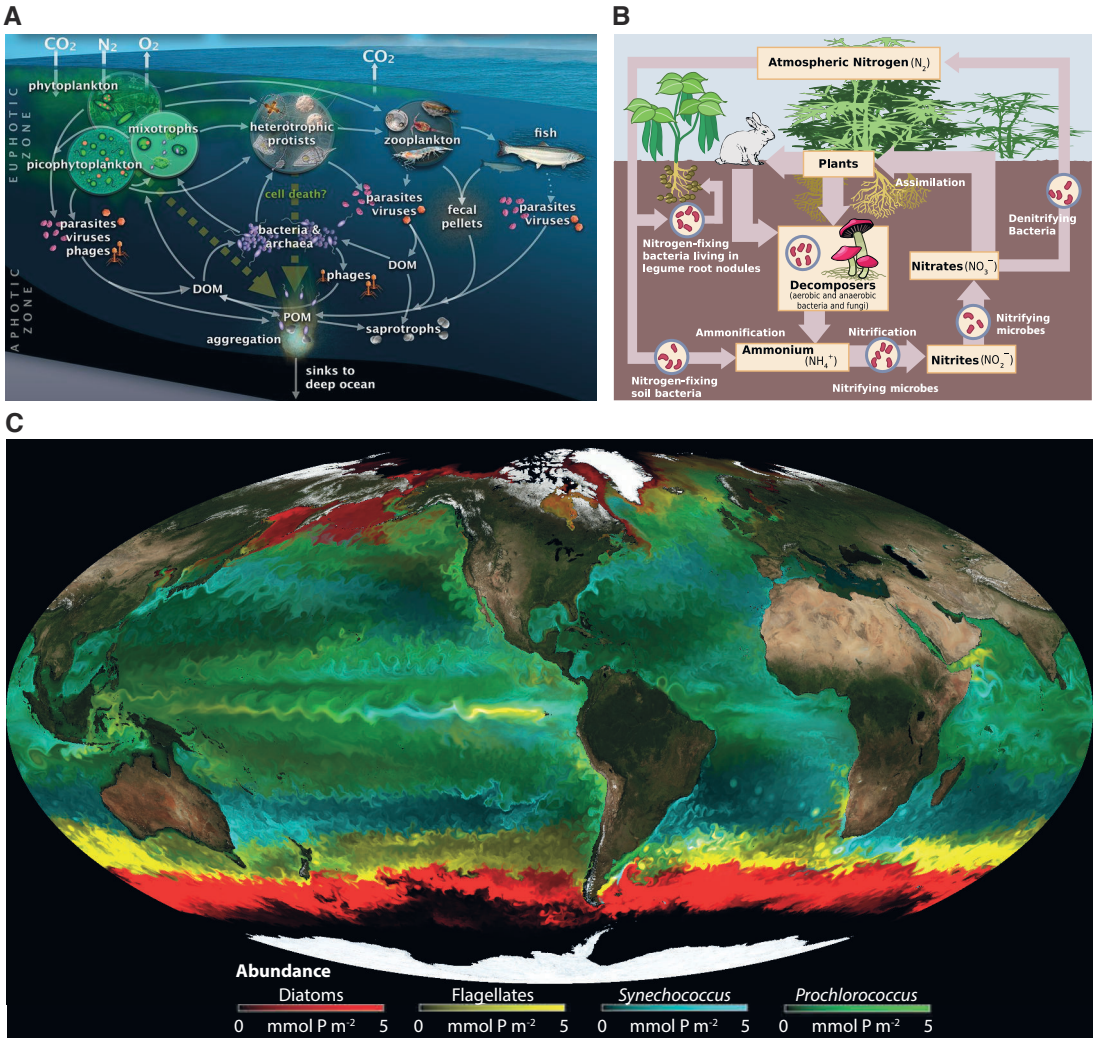


Figure 1.1: The role of microorganisms in global biogeochemical cycles and as primary producers. **A** Microbial phytoplankton absorbs CO_2 from the atmosphere to create complex organic compounds, which are in turn consumed by heterotrophic or mixotrophic microbes. The CO_2 produced by these species, as well as by higher-order heterotrophic organisms that feed on them, subsequently gets recycled into the atmosphere, thus completing the carbon cycle. Adapted with permission from Worden et al. (2015). **B** Nitrogen-fixing microbes perform essential steps in the global nitrogen cycle by converting molecular nitrogen (N_2) into bio-available Ammonium (NH_4^+), which is either consumed directly by plants (via root nodules) or further oxidized to nitrate (NO_3^-) by nitrifying microbes. Nitrate is subsequently assimilated by plants (and, through plant consumption, also by animals) or converted to N_2 by denitrifying microbes, re-entering the atmosphere. Adapted with permission from Wikimedia Commons (2009). **C** Global distribution of the most important marine primary producers, *Prochlorococcus* and *Synechococcus*, which are among the most abundant and successful microbial groups on Earth. Adapted from a graphic kindly provided by Prof. Dr. Michael Follows (The Darwin Project, Massachusetts Institute of Technology).

estimated to be the most numerous group of photosynthetic organisms on the planet and constituting the most important primary producer in Earth's oceans (Flombaum et al. (2013); see Figure 1.1 C). More exotically, extremophile chemosynthetic microbes are essential primary producers in the deep-sea, associated with hydrothermal vents, seeps or whale and wood falls, where whole (and surprisingly diverse) ecosystems rely on their ability to generate energy-rich organic molecules from minerals and other inorganic chemical compounds, provided for instance by vents, without the use of sunlight (Dubilier et al. (2008); see Figure 1.2 B).

The extraordinary metabolic potential of microbes is also helpful in degrading a wide

array of potentially toxic compounds in the environment, which re-inserts elements into their corresponding biogeochemical cycles and thus makes them available to other organisms. A prime example is the removal of surplus NH_3 through the process of nitrification (Madigan et al., 2014), now known to be predominantly mediated by archaea (Hatzenpichler, 2012), which is a pivotal step in the nitrogen cycle (see Figure 1.1 B). Other examples include the degradation of contaminating hydrocarbons from oil spills (Das and Chandran, 2011), xenobiotics such as polyethylene (Nowak et al., 2011), pesticides (Kumar et al., 1996) and pharmaceuticals (Benotti and Brownawell, 2009).

Apart from these broad effects on life on Earth, countless well-described specific symbiotic relationships between multicellular and unicellular organisms exist (see Figure 1.2), covering all branches of the tree of life McFall-Ngai (2008). One of the most impactful ones is the tight mutualistic symbiosis between nodulated plants, most notably legumes such as *Glycine max* (soybean) and *Medicago sativa* (alfalfa), and microbial species from the *Alpha-* and *Betaproteobacteria*, collectively called rhizobia (Madigan et al., 2014). Rhizobial strains infect specific host species and initiate the well-studied formation of root nodules, structures in the root where these strains accumulate to perform nitrogen fixation under favorable nutrient and oxygen conditions (Desbrosses and Stougaard (2011); see Figure 1.2 C).

In addition to rhizobial bacteria, an estimated 80% of land plants also harbor essential mycorrhizal fungal symbionts, which provide the host plant with water and nutrients, such as phosphate, in exchange for carbon (Deacon, 2013; Parniske, 2008). These fungal symbionts can have profound effects on plant diversity, community structure and productivity, with some combinations of fungal species increasing biomass yields by several folds in certain agricultural host plant species (Heijden et al., 1998). Intricate signaling pathways have been described in plants to sense and recruit suitable bacterial and fungal symbionts (Desbrosses and Stougaard, 2011).

A widespread class of mutualistic relationships with animals is conveyed by bioluminescence, which provides distinct advantages to host animals in low-light environments, most notably marine habitats, with regards to mating, detection and attraction of prey, and camouflage (Widder, 2010). A remarkable example of the latter is the squid *Euprymna scolopes*, which begins life symbiont-free and then picks up specimen of the specific co-evolved gammaproteobacterium *Vibrio fisheri* (see Figure 1.2 A). These microorganisms subsequently become enriched in a specialized symbiotic light organ and provide the squid with counter-illumination, mimicking the moon and starlight coming from the upper water columns to impede detection by predators (McFall-Ngai and Ruby, 1998). In exchange, *V. fisheri* is provided with oxygen, nutrients, and a safe habitat by its host. Strikingly, *V. fisheri* possesses sensor genes for host specificity and furthermore strongly outcompetes non-bioluminescent

mutants, as well as other microbial groups from closely related squid species, within the host (Mandel et al., 2009; McFall-Ngai and Ruby, 1998). This relationship is thus a remarkable example for an intimate co-evolutionary history between microbial symbiont and host.

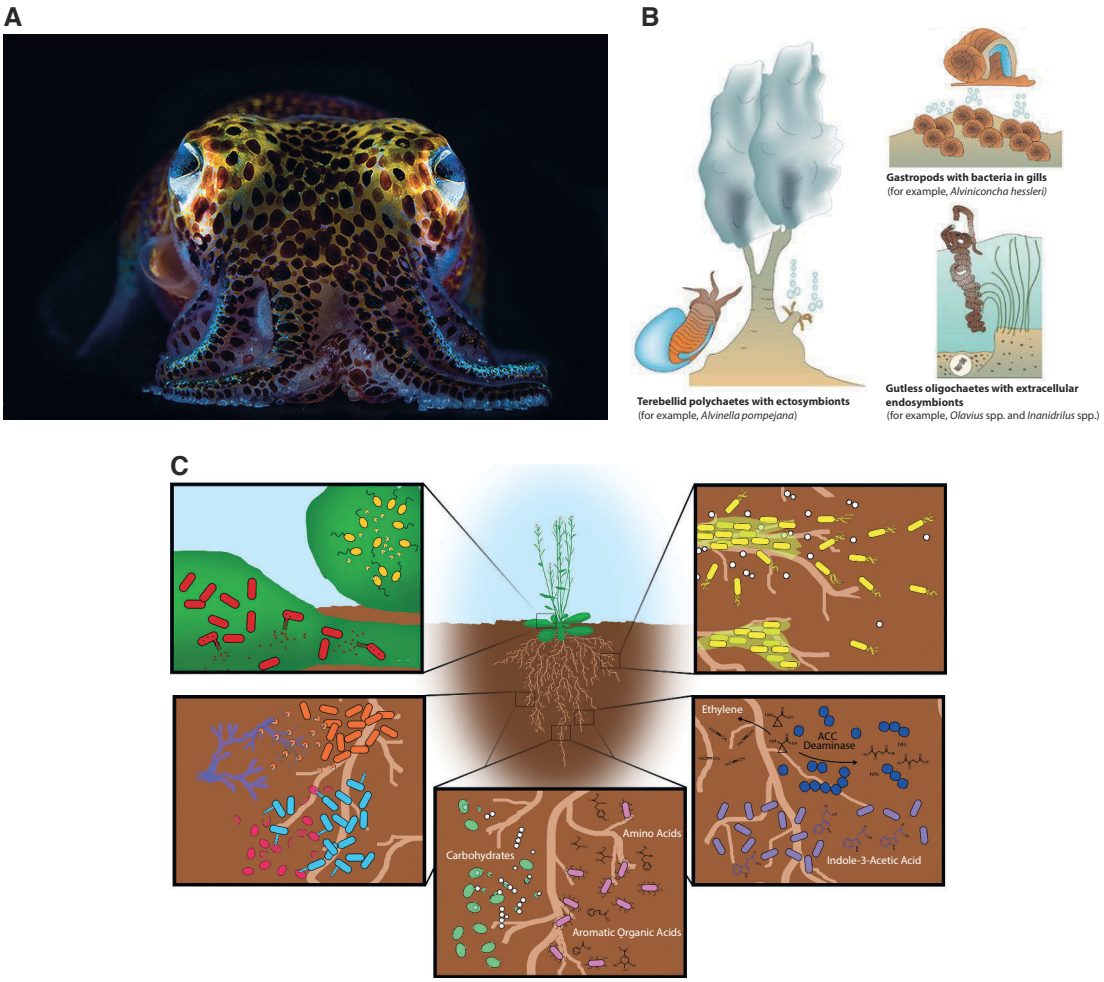


Figure 1.2: Important symbioses between microbial symbionts and animal or plant hosts. **A** The squid *Euprymna scolopes* developed a tight symbiotic relationship with *Vibrio fisheri*, which aggregate in a specific light organ within the squid's mantle and provide it with bioluminescence. Credit due to Mattias Ormestad. **B** Chemosynthetic microbes are essential primary producers in the deep-sea. In this environment, they are partners in numerous symbiotic relationships, for instance as ectosymbionts on the surface and gills of diverse animals, or as endosymbionts within gutless worms. Graphic adapted with permission from Dubilier et al. (2008). **C** Microbes have various interactions with plant hosts, for instance mediated by the production of important nutrients and hormones, or through defense against pathogens. Colonization is driven by multiple factors, such as chemotaxis and adherence factors which promote biofilm formation. Graphic adapted with permission from Levy et al. (2018).

A more complex example of animal-microbe mutualism are corals, which are dependent on single-celled eukaryotic dinoflagellates (*Zooxanthellae*) for their survival. These symbionts perform photosynthesis to generate energy and nutrients for themselves and their host and are in turn provided with ammonia and phosphate in an environment starved for these essential inorganic nutrients (Hoegh-Guldberg, 1999; Stat et al., 2008). Both the animal host and the zooxanthellae symbionts furthermore depend on nitrogen-fixing bacterial symbionts—mostly cyanobacteria, but also alphaproteobacte-

ria from the order *Rhizobiales*[2]—to obtain bioavailable nitrogen (Lema et al., 2012; Lesser et al., 2004). Triggered by stress, most prominently global warming-related heat stress, catastrophic events called coral bleaches can occur, in which the animal hosts eject their zooxanthellae symbionts and transition into a vulnerable state with high mortality rates, resulting in the destruction of whole ecosystems (Hoegh-Guldberg, 1999).

Other marine examples of animal-microbe mutualism include the previously mentioned symbiotic relationships between animals and chemosynthetic microbes in the deep-sea. To current knowledge, these interactions stretch across seven animal phyla (Dubilier et al., 2008) and include curious hosts, such as gutless worms, which are obligately reliant on their symbionts (Kleiner et al. (2012); see Figure 1.2 B). Up to 40% of the biomass of some sponge species furthermore consists of microbial symbionts (Friedrich et al., 2001).

Intriguingly, the possibility of microbial endosymbiont exchange promoting the evolution of certain types of animal behavior, for instance coprophagy and sociality, has been proposed (Lombardo, 2008), extending earlier work on herbivore evolution (Troyer, 1984). Microbes can furthermore confer distinct advantages to biological pests invading a new habitat, as was shown for instance in invading insects (Lu et al., 2016; Vilcinskas et al., 2013). Additionally, there is evidence for the tight control of endosymbionts in insects via a controlled administration of antimicrobial peptides by the host (Login et al., 2011).

Especially the close relationships between vertebrate hosts and their microbial symbionts have been studied in remarkable detail in recent years, unveiling for instance a number of intricate connections between symbionts and organ homeostasis (see Figure 1.3). Studies have focussed in particular on the body habitat typically harboring the vast majority of microorganisms: the gastrointestinal tract. Close co-diversification between a wide variety of vertebrate hosts and their gut microbiota has been observed (Ley et al., 2008; Sharpton, 2018) and gut communities have furthermore been found to generally cluster by host phylogeny, albeit the picture is complicated by digestive tract physiology (hindgut vs. foregut fermenters) and environmental factors, such as host diet (carnivorous, omnivorous, herbivorous) (Muegge et al., 2011; Sharpton, 2018). Some hosts, for instance ruminant species, harbor specialized organs which accommodate highly diverse microbiomes and are specialized on the digestion of complex carbohydrates, normally inaccessible for mammal nutrition. This intimate relationship between host and microbiome enabled the extraordinary success story of herbivory, which emerged from within carnivorous mammals[3] and is now estimated to cover up to 80% of known mammals (Chapman, 1997; Ley et al., 2008). This example

[2] similar to many terrestrial plants, see above

[3] Curious parallels furthermore exist between gut microbiomes of modern terrestrial herbivorous mammals and carnivorous cetaceans (whales and dolphins), which both evolved from herbivorous ancestors (Sanders et al., 2015).

highlights the immense fitness advantages that host-microbe interactions can confer and underlines the important role of coevolution in supporting an efficient exploration of the fitness landscape (Hutzil et al., 2018; Solé and Sardanyés, 2014).

Competition experiments, in which *Lactobacillus reuteri* strains isolated from several non-murine vertebrate host species were transplanted into mouse hosts, showed high specificity and distinct competitive advantages of rodent-derived strains over strains from non-murine host species (Oh et al., 2010). Similarly, in reciprocal transplant experiments between germ-free zebrafish and mice, transplanted communities shifted abundance profiles to mirror those of closely related native species, indicating strong selective pressures by the host environment (Rawls et al., 2006). Notably, transcriptional host responses to microbiome transplants were still highly conserved between host species in these experiments, indicating evolutionarily relevant host-microbe interactions. But also for more closely related hosts, specific microbiome requirements are becoming apparent: for instance, reduced survival rates and overall fitness of hosts from a particular murine species were observed following transplantation of microbiota from other murine species (Brooks et al., 2016). Host genetics may partly explain this pattern, as specific quantitative trait loci in host genomes were found to modulate single microbial taxa, but also groups of closely and even distantly related taxa (Benson et al., 2010).

Gut-residing microbes were traditionally regarded as commensals, which benefit from the nutrient-rich and relatively predation-free environment but have no effect on their host ("eating at the same table", McFall-Ngai (2008)). However, modern microbiome research is now uncovering the mutualistic nature of host-microbiome relationships from various angles. A prime example is the importance of diverse microbial clades for the digestion of complex compounds. Important representatives of these endosymbionts are species from the genera *Bacteroides* (including the well-studied *Bacteroides thetaiotaomicron*) and *Bifidobacterium* (Belenguer et al., 2006; Martens et al., 2009), which process dietary polysaccharides and proteins to produce compounds that nourish the host and other beneficial members of the microbiome (the latter via cross-feeding) (Flint et al., 2008; Hooper et al., 2002). Even single microbial species have been observed to noticeably modulate host inflammatory responses, as for example *Faecalibacterium prausnitzii* (Lopez-Siles et al., 2017), and microbial activity has been linked to the modulation of energy storage and harvest within the host (Martin et al., 2007).

Many health-associated metabolites appear to be of microbial origin and are for instance related to the body mass index (BMI) and coronary heart disease, which further emphasizes the importance of the microbiome for host health (Holmes et al., 2008). Vital microbial products include for instance essential vitamins and amino acids, but also short-chain fatty acids (SCFAs), which are recently being studied in

particular depth. The most important SCFAs are butyrate, acetate and propionate, typically produced in complex food chains of cross-feeding microbial species (Belenguer et al., 2006; Falony et al., 2006). SCFAs are important energy providers for the colonic epithelium that promote energy homeostasis and result in various health benefits (Byrne et al., 2015). Similar to other microbial metabolites, SCFAs can be sensed by the host via specific sensory receptors, which triggers specific host responses (Bäckhed, 2012; Samuel et al., 2008). This intricate mechanism constitutes another prime example for the concerted interplay between microbial endosymbionts and their host.

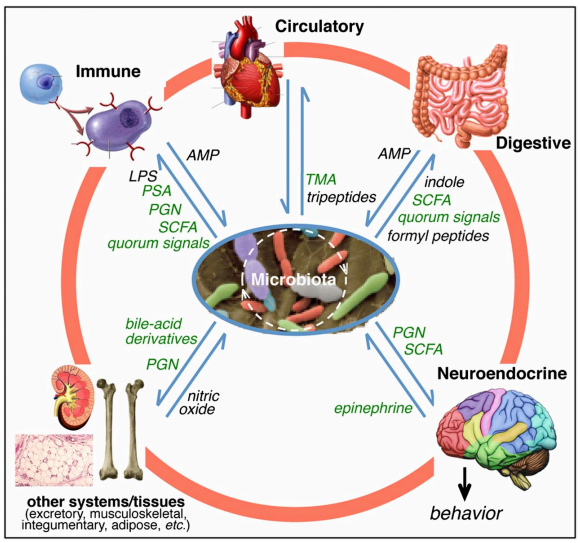


Figure 1.3: The diverse roles of microbial endosymbionts in animal hosts. Microorganisms can affect every major organ of the host body, with pronounced implications for homeostasis. While effects on the digestive tract and immune systems have been extensively studied, interactions with the nervous system and the impact on host behavior have more recently become a focal point. Credit due to McFall-Ngai et al. (2013).

Another important class of host-microbe relationships are close interactions between intestinal microbiome and the host immune system (see Figure 1.3). While microbial mutualists confer immense advantages to their host (as described in the previous paragraphs), sophisticated control mechanisms are necessary to allow for the specific selection of beneficial and exclusion of harmful species. Colonic antimicrobial defenses and modulatory mechanisms are therefore regulated by a complex machinery[4] (Mukherjee et al., 2008). Indeed, immune defenses have to balance a careful trade-off between a tolerant state and tighter control. While a more tolerant immune system is more open to beneficial microbes and maximally secures their benefits (while also avoiding autoimmunity), more strictness helps to keep pathogens out and to prevent harmful invasions of colonic tissues by commensal microbes, which can turn rogue in the wrong place (Abbeele et al., 2011).

An important separation within the intestinal microbiome lies between luminal and

[4]This realization lead to the proposal that increased endosymbiont control may even be one of the major driving forces behind the evolution of the vertebrate adaptive immune system (McFall-Ngai, 2007).

mucosal communities. Luminal communities reside inside the gut lumen and mainly play a nutritional role, only indirectly interacting with the colonic epithelium (via diffusing metabolic products) and having limited to no access to the immune-molecule-rich mucosal layers of the colon. Mucosal mutualists, on the other hand, inhabit the mucosal layers and show a markedly different community structure compared to luminal microbes (Zoetendal et al., 2002). Mucosal microbes furthermore are tolerated by the abundant mucosal defense molecules produced by the host and can survive the increasing oxygen gradient within the mucus, while feeding on host-provided mucins (Abbeele et al., 2011; Meyer-Hoffert et al., 2008). Mucosal communities play important roles in protecting the colonic epithelial tissue from pathogens: they decrease the mucosal pH, produce specific antimicrobials, and compete with pathogens for nutrients and cohesion spots on the epithelial surface, while also activating and modulating the host defenses for pathogen recognition and tolerance of commensals (Abbeele et al., 2011; Hooper et al., 2003; Mazmanian et al., 2005; Vaishnava et al., 2008; Willing et al., 2009).

Recently, specific microbiome patterns have been associated with human diseases not traditionally perceived as microbe-related disorders, including but not limited to obesity, inflammatory bowel disease, Crohn's disease, colorectal cancer and autism (Adams et al., 2011; Beaumont et al., 2016; Nishida et al., 2018; Willing et al., 2009; Zeller et al., 2014). The modern sterile western lifestyle has furthermore been implied in an increased risk for autoimmune diseases, allergies and inflammatory gastrointestinal diseases due to lack of exposure to important microbial pathogens and commensals, resulting in the incomplete tuning of the immune system during development (the "hygiene hypothesis", Filippo et al. (2010) and Rook (2007)). Moreover, specific pre- and probiotics have been proposed to provide various health benefits, for instance by combating specific diseases (Cani and Vos, 2017; Hemarajata and Versalovic, 2013; Slavin and Slavin, 2013; Tankou et al., 2018).

However, pinpointing single causative species has not been successful (yet) in many cases[5] and causality, mechanisms and generality of disease induction and probiotic health improvement require more research (Ma et al., 2018; McKenney and Pamer, 2015; Schmidt et al., 2018). Alarmingly, newer results suggest that effects of probiotic treatment may be highly host-dependent and can even include increased disease susceptibility (Quin et al., 2018; Suez et al., 2018; Zmora et al., 2018). Similarly, care is advised when interpreting observed disease-microbiome associations, as illustrated strikingly in the case of type 2 diabetes, for which pronounced microbiome shifts were initially reported to be directly associated to the disease (Karlsson et al., 2013; Qin et al., 2012), but later shown to be confounded by the anti-diabetic drug metformin (Forslund et al., 2015).

[5] possibly due to a polymicrobial nature of diseases (Peters et al., 2012)

A more recent line of research suggests an intense bilateral interplay between the brain and intestinal microbes, conveyed by the so-called "gut-brain-axis", through which intestinal microbes impact neuronal activity and behavior of the host and vice-versa (Carabotti et al., 2015; Johnson and Foster, 2018; see Figure 1.3). One mediator of this relationship are microbe-derived neuroactive compounds (Lyte, 2013), which may influence cognitive patterns, such as affect and motivation, and potentially play a role in neurological diseases (Mayer, 2011). Some effects have been traced to single microbial species: *Lactobacillus* and *Bifidobacterium* species have for instance been observed to reduce anxiety and depression-like symptoms in both mice and humans (Johnson and Foster, 2018).

Apart from microbe-host interactions, biological interactions between different microbial species play a fundamental role in nature. All important interaction classes used to describe the ecology of multicellular organisms are also observed between microbes, including cooperative relationships (mutualism and commensalism), as well as antagonistic relationships (competition, predation, parasitism and amensalism) (Lidicker (1979); see Figure 1.4 A). As I detail below, all of these can have profound effects on ecosystem function.

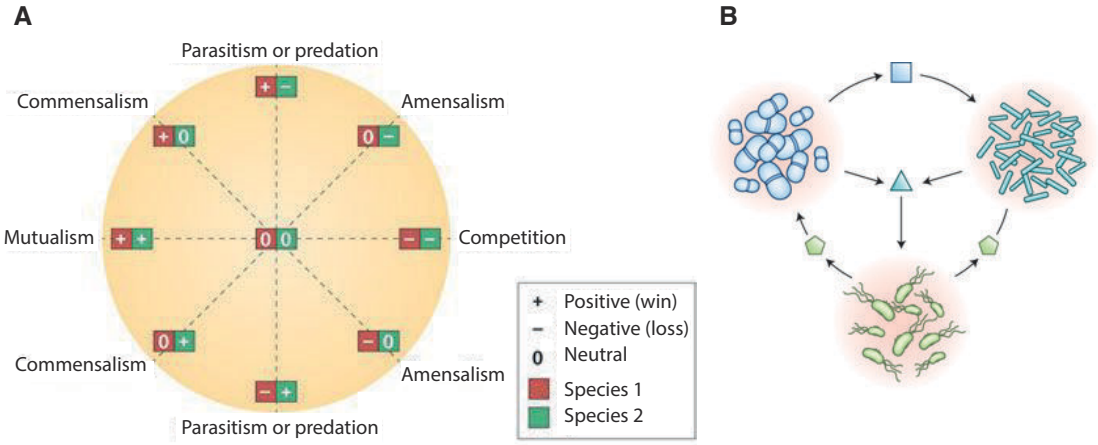


Figure 1.4: Ecological interactions between microbes. **A** All types of known ecological relationships have been observed between microbes. These interaction types are defined based on patterns of benefit, harm and neutrality between interaction partners. Credit due to Faust and Raes (2012). **B** One example of interactions are cross-feeding relationships, for instance commensalism and complex mutualistic feedback loops. Credit due to Tang (2019).

Mutualism, a cooperative relationship in which both partners benefit, can lead to tight and typically long-term interactions (Wingreen and Levin, 2006). This ecological interaction type can for instance be mediated through intricate metabolic dependencies, an example of which is the concerted interplay of diverse metabolically interdependent microbial partners driving the global carbon and nitrogen cycles in a spatially and temporally separated manner (Falkowski et al., 2008). Similarly, complex food chains of mutualistic microbes exist in the vertebrate gut (see Figure 1.4 B), featuring extensive metabolic complementarity and cross-feeding, where primary fermenters such as *Bacteroides* can be viewed as gateway metabolizers through which carbohydrates enter

the food web (Clavel et al., 2006; Fischbach and Sonnenburg, 2011). This also includes the degradation of waste products, as for instance in the production of SCFAs, where H_2 consuming microbes (e.g. methanogens) metabolize the H_2 produced by primary fermenting microbes, leading to a reduction of the environmental H_2. This in turn helps their fermenting partners to maintain metabolic efficiency[6] (Macfarlane and Macfarlane, 2003).

Intriguingly, mutualism in the human gut appears to be so important that the majority of antibiotic-related disturbance effects may be attributable to their impact on cross-feeding (Dethlefsen et al., 2008). It is in fact such a fundamental trait that it can even be evolved under laboratory conditions: an experimental evolution study performed by William Harcombe resulted in a mutualistic relationship between *Escherichia coli* and *Salmonella enterica* strains, where the *S. enterica* strain, which started out metabolizing waste products of the *E. coli* strain, subsequently evolved to also provide an essential amino acid in return, completing the cooperative loop (Harcombe, 2010). Similarly, an intimate mutualistic interaction between two strains from the soil-dwelling species *Pseudomonas putida* and *Acinetobacter* sp. was experimentally evolved within a biofilm flow chamber (Hansen et al., 2007).

Apart from metabolic dependencies, mutualistic relationships can also occur for defensive purposes. A textbook example of this are biofilms, i.e. microbial communities dwelling on surfaces and surrounded by extracellular polymeric substances. These surrounding compounds are typically produced by specific members as a shared resource, protecting the community as a whole from toxic substances (antibiotics) and antagonists, while also allowing more efficient and localized communication, nutrient sharing and division of labour. Biofilms may in fact be the typical mode of existence in many bacteria (Kolter and Greenberg, 2006) and show intriguing similarities to multicellular organisms, in some cases displaying behavior resembling that of a single entity (Kaiser and Losick, 1993; Shapiro, 1998). Essential for the accumulation and maintenance of biofilms is inter-microbial communication (Scherlach and Hertweck, 2018), as realized for instance through the well-studied phenomenon of quorum sensing (Mashburn and Whiteley, 2005; Whiteley et al., 1999). This mechanism leads to concerted interspecific and community-wide reactions, such as swarming or the production of defense molecules, in response to population density and external signals (Waters and Bassler, 2005). An interesting variation of biofilm formation was recently observed for an uncultured *Euryarchaeon*, which uses physical extrusions to simultaneously "grapple" and hold onto multiple other archaea or bacteria, forming a regular multi-species matrix (Perras et al., 2014).

Microbial mutualism may be the main driver for the emergence of complexity, in

[6]Computational predictions furthermore suggest that these metabolic complementarity-driven mutualistic relationships may be widespread across diverse habitats (Zelezniak et al., 2015).

particular the evolution of the eukaryotic cell and multicellularity. It was for instance proposed that energetic advantages provided by an alphaproteobacterial endosymbiont (the protomitochondrion) to its putative primordial archaeal host reduced energetic constraints of the host. This subsequently may have triggered genome expansion and the evolution of increasingly complex features, eventually resulting in the first bonafide eukaryotic cell (Lane and Martin, 2010; Martijn and Ettema, 2013). Similarly, certain modes of beneficial coexistence, as for instance biofilms, may have facilitated the cooperation of multiple cell types (species) in close proximity over extended periods of time, leading to an increasing interdependence and thus providing a stepping stone towards true multicellularity (Lyons and Kolter, 2015; West et al., 2006).

In the second cooperative interaction type, commensalism, one interaction partner benefits while the other neither benefits nor is harmed (i.e. it experiences no net effect). Albeit more typical for scenarios in which the host is orders of magnitude larger than the symbiont (e.g. commensal microbes of plant or animal hosts), commensalism can also be found between microbial species (Bertrand et al., 2015). For instance, previously discussed primary producers, such as photosynthetic and chemosynthetic microbes, can be partners in commensal relationships if they provide side products not used by themselves, which can then be metabolized by partner microbes. More subtle but nonetheless important forms of commensalism include also the sequestration of molecules that make environmental resources available for metabolization by other microbes, for instance siderophores in iron sequestration (Griffin et al., 2004). However, this relationship may turn into antagonism if the activated resource is limited or mutualism if the non-producers provide reciprocal benefits (as typical in biofilms). Another well-studied example are fermenting microbes utilizing lactate during the ripening process of dairy products, which increases the pH in the product and paves the ground for less acid-tolerant successional microbes (Irlinger and Mounier, 2009; Mounier et al., 2008). Similarly, aerobic microbes can create anoxic environments in the human oral cavity, preparing a habitat for oxygen-sensitive colonizers (Mark Welch et al., 2016). Commensalism is also commonly found in biodegradation, for instance cellulose degradation, where cross-feeding relationships abound (Leschine, 1995).

The second broad class of ecological interactions—antagonistic relationships, in which at least one interaction partner is harmed—is also commonly observed between microbes (Feichtmayer et al., 2017). One widespread and important class of antagonistic interactions is the competition for shared resources in overlapping niches, which leads to disadvantages for both competitors (Hibbing et al., 2010). Typically, competitive advantages such as a more efficient metabolism and higher growth rates will give one species the edge, allowing it to outgrow its competitor and ultimately eliminate it from the niche (competitive exclusion, Bauer et al. (2018)). To circumvent this, the disadvantaged species usually attempts to reduce the competitive pressure

by slightly changing strategies, which results in niche partitioning. This may involve switches in metabolic requirements (i.e. the metabolization of compounds not usable by the other species), the employment of defense strategies, such as the production of toxins and antimicrobials, or the change of its general survival strategy, for instance by putting more emphasis on motility and dispersal to colonize newly opening niches more quickly than it's metabolically more efficient competitor (Bauer et al., 2018; Hibbing et al., 2010).

However, such evasion strategies can be exploited, as is displayed by *Eubacterium rectale* in the mouse gut. *E. rectale* competes with *B. thetaiotaomicron* for CO_2 (provided primarily by the host) and can sense the presence of its competitor. After detecting *B. thetaiotaomicron*, *E. rectale* upregulates PEP carboxykinase which allows it to more efficiently deplete CO_2 than its antagonist. *B. thetaiotaomicron* is forced to respond by switching from propionate to the less CO_2-intense production of acetate, which in turn can be metabolized by *E. rectale*, thereby effectively killing two birds with one stone (Fischbach and Sonnenburg, 2011; Mahowald et al., 2009).

Of particular importance for host health is colonization resistance, a competitive phenomenon in which beneficial microbial species block available intestinal niches from pathogens, thereby preventing invasion by these harmful groups. This protective effect was for instance observed as a defense against *Clostridium difficile*, *Salmonella enterica* and *Shigella* (Buffie et al., 2015; Lawley et al., 2008; Pérez-Cobas et al., 2015; Stecher et al., 2010), and can be weakened or lost in the presence of intestinal dysbiosis (Buffie and Pamer, 2013).

Another type of antagonistic relationships is predation, in which one interaction partner (the predator) benefits from harming the other (the prey) through consumption. This interaction is prominently observed between predatory unicellular eukaryotes, classically protozoa or different amoebae species, and bacterial prey. When grazing on bacteria, the predators utilize their size and advanced, flexible cytoskeleton to engulf their prey, followed by digestion (Madigan et al., 2014). However, also bacteriovorus bacteria have been identified, which employ a variety of feeding strategies, from invasion of the prey's periplasm (*Bdellovibrio bacteriovorus*), with some parallels to parasitic behavior, to extracellular attachment followed by introduction of hydrolytic enzymes into the prey cell (*Vampirococcus* sp.) (Feichtmayer et al., 2017; Guerrero et al., 1986).

Parasitism is another common antagonistic relationship, in which one partner benefits, while the other is harmed without acute risk of death. A large diversity of microbes have been described as parasites of other microbes: for instance cyanobacteria, which often form massive blooms and are therefore prime targets for predators, but also parasites, host a wide variety of unicellular fungal, protist and bacterial parasites (Haraldsson et al., 2018). Further examples include parasites of microbial parasites

(so-called "hyperparasites"), which appear to be common in nature (Parratt and Laine, 2016), and microbial "cheaters", which benefit from and deplete commodity goods provided by other microbes without providing benefits in return, as observed for instance for iron sequestration (West and Buckling, 2003) (previously mentioned also in the context of commensalism). An exotic candidate for parasitism is the archaeon *Nanoarchaeum equitans*, which has been suggested to use other archaea from the genus *Ignococcus* as host (Waters et al., 2003), possibly making this species the first described archaeal parasite.

The final class of antagonistic relationships is amensalism, in which one partner is harmed without any effect on the other partner. Well-known examples are the production of antibiotics, for instance the classic case of penicillin produced by several species from the fungal genus *Penicillium*, which are secreted to weaken or kill competitors and predators, but usually also induce collateral damage to non-threatening microbes (Lemos et al., 1991; Long et al., 2013). Another example was observed in mixed cultures between *Lactobacillus casei* and *Pseudomonas taetrolens*, in which *P. taetrolens* showed highly unstable growth and viability profiles, linked to the lactobionic acid produced by *L. casei*, while *L. casei* appeared unaffected in the co-culture (García et al., 2017).

Interestingly, the previously described interaction types can be highly dependent on both biotic and abiotic factors, and may thereby change from one type to another when conditions are altered (Buffie et al., 2015; Liu et al., 2017). For instance, mutualistic interactions can transition to competition or parasitism in benign environments where mutual cooperation is not required (Hoek et al., 2016). A health-related example is the observed protective effect of *Clostridium scindens* against *Clostridium difficile* overgrowth, which importantly is only realized in the presence of primary bile acids, which *C. scindens* metabolizes into protective secondary bile acids (Buffie et al., 2015). Similarly, the inclusion of new species in a community can shift the metabolic behavior and active pathways of resident members, for instance through niche partitioning. An example of this are metabolic switches in *B. thetaiotaomicron*, triggered by the addition of *Bifidobacterium longum* or *E. rectale*[7] (Mahowald et al., 2009; Sonnenburg et al., 2006), which in turn changes interactions with other members of the intestinal microbiome. A similar example is the increased production of bile acids by intestinal commensals in response to prebiotic food intake, which can render the intestinal environment unfavorable for pathogens like *Shigella* (Alvarez-Acosta et al., 2009; Rabbani et al., 2009).

The situation is further complicated by multiple interaction types potentially being active in parallel. An example is lateral gene transfer (LGT), in which the comprehensive gene pool of a community is treated as a shared resource (Woese, 2002). LGT is more

[7] as discussed earlier

frequent between closely related strains and species (Soucy et al., 2015), making it a possible mechanism of kinship selection and more common between mutualists (Strassmann et al., 2011). Nonetheless, shared genes may also confer advantages to competitors and predators, blurring the picture of cooperation and antagonism.

1.1.2 Fundamental ecosystem properties

> *When one tugs at a single thing in nature, he finds it attached to the rest of the world.*
>
> —John Muir

Ecosystems and their properties have been studied theoretically and empirically for a long time—for instance, the first analysis of a food web was presented as early as 1912 (Pierce et al. (1912); see Figure 1.5 A). Even earlier, Charles Darwin acknowledged the importance of the complex intertwined relationships between all organisms through his "tangled bank" metaphor (Darwin, 1859). Traditionally, ecological network analysis focussed mainly on ecological systems of animals and plants, including for instance food-webs (Pimm et al., 1991), mutualistic networks of plants and pollinators or seed dispersers (Bascompte and Jordano, 2007; Nilsson, 1988) and host-parasite networks (Askew, 1961) (see Figure 1.5 B). Furthermore, classic network ecology studies were mainly concerned with the analysis of simple pairwise interactions, with little consideration of emergent network properties (Bascompte, 2009).

Recently, graph- and network-theoretical approaches have become a major driving force for fundamental insights into many types of complex networks, including engineered networks, such as power grids or the internet, but also social networks, systems in statistical physics, and biological networks. The latter cover for instance nervous systems and metabolic networks (Albert and Barabasi, 2002; Newman, 2003; Strogatz, 2001). Mirroring this trend in other fields, a more holistic, network-centric perspective with focus on emerging large-scale patterns is also gaining traction in ecology. In consequence, network-theoretic approaches are being successfully applied to a growing number of ecosystem models (Bascompte, 2009; Montoya et al., 2006). These studies typically investigate and interpret general graph properties, such as measurements of community structure (including modules, nestedness and compartmentalization), transitivity, connectance, average shortest paths, betweenness and specific network motifs[8] (Bascompte, 2009; Delmas et al., 2019; Montoya et al., 2006; Proulx et al., 2005; see Figure 1.6 B for selected examples). General graph characteristics have enabled the study of a number of higher-order properties, such as network robustness, stability and resilience (Albert et al., 2000; May, 2001; Oliver

[8] which furthermore inform network alignment methods

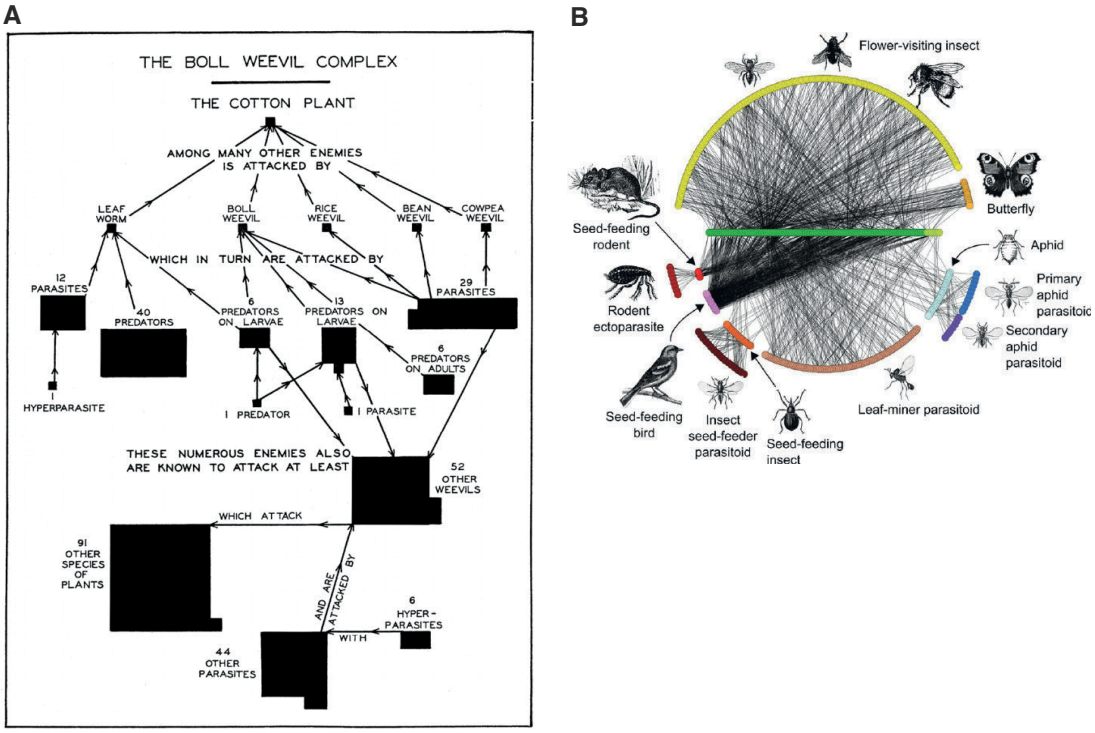


Figure 1.5: Ecological networks of plants and animals. A Possibly the first depiction of an ecological food web, described by Pierce, Cushan and Hood in 1912 (Pierce et al., 1912). **B** Example of an ecological network with complex interactions between plants and various mutualistic (pollinators, seed dispersers) or parasitic animal partners. Credit due to Bohan et al. (2013).

et al., 2015; Pimm, 1984), as well as the classification of networks into fundamental graph-theoretical classes, most notably small-world and scale-free networks (Albert and Barabasi, 2002; Newman, 2003; Strogatz, 2001; see Figure 1.6 A).

In order to obtain networks to which these methods can be applied, pairwise interactions have to be identified for all species within the network (optimally including signs and strengths, which may be difficult to quantify in practice). For animal-plant systems, interactions are typically inferred through empirical observations (aided by technical equipment, such as camera traps), for instance on the types and frequencies of pollinating animals visiting different plant species, prey species being caught by particular predators, or different seed types being found in feces of specific animal species (Pellissier et al., 2018).

Ecological studies with a graph-theoretical perspective have yielded insights into a number of interesting ecosystem properties (see Figure 1.6). For instance, many ecological networks across geography, habitat types and network classes (e.g. food webs, mutualistic networks), have been shown to be densely connected and complex (Montoya et al., 2006), with properties of both random and regular networks (Newman, 2003). This has implications for species diversity, which tends to increase with higher numbers of direct and indirect interactions between species (Montoya et al., 2006).

In addition, most ecological networks exhibit the small-world property, which means that average shortest distances (i.e. the "diameter" of the network) between species

are small[9] (Amaral et al. (2000); see Figure 1.6 A). Small-world networks furthermore have high clustering coefficients compared to random networks, which implies that one species can be reached from others within a small number of steps and that neighbors of a species are typically also connected (Montoya and Solé, 2002; Williams et al., 2002). While the small-world property may enable quicker responses to perturbations, increasing system homeostasis (Montoya and Solé, 2002), it paradoxically also implies that negative effects can more quickly propagate through the whole network (May, 1972), making it less robust to perturbations. However, empirical ecosystems generally display robustness and resilience, which is why non-random properties of ecological networks have been suggested to explain this conundrum (dissected in the "diversity-stability debate", Jacquet et al. (2016) and McCann (2000)).

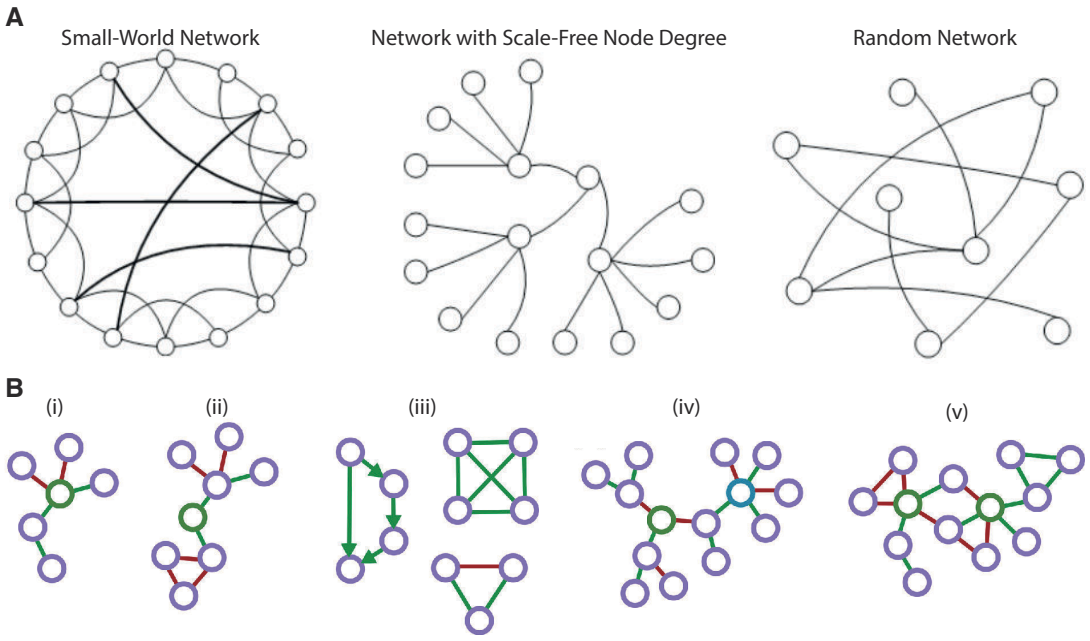


Figure 1.6: A selection of general properties found in complex networks. A Different network classes are distinguished depending on node degree distributions and connectivity patterns, such as the network diameter. Adapted with permission from Huang et al. (2005). **B** Important nodes within ecological networks include hubs (i) and nodes with high betweenness centrality (ii). Networks can furthermore contain specific motifs (with potential functional relevance) (iii), be assortative (iv) and show fragility to the removal of specific nodes (v). Adapted with permission from Röttjers and Faust (2018), where further details and explanations can be found.

Another widespread property of ecological networks is their right-tail heavy degree distribution, which means that most species have a low number of interaction partners, while a few have disproportionately many. This property is in sharp contrast to random networks, where node degrees follow a Poisson distribution (Erdos and Rényi, 1960). The right-tail heavy property is reminiscent of networks with a scale-free node degree distribution (see Figure 1.6 A), in which the degree distribution follows a power-law, explainable for instance by a "the rich-get-richer" (or preferential attachment) effect (Barabási and Albert, 1999). Examples of such networks include the World Wide Web,

[9] or, more formally, the diameter scales logarithmically with the number of nodes (Amaral et al., 2000)

networks of scientific collaborations and citations, protein interaction networks and a social network of sexual contacts (Newman, 2003)[10].

In contrast, ecological networks usually do not fit a power law model most closely. Instead, they are typically better explained by a *truncated* power law model, in which the maximum neighborhood size is bounded (Amaral et al., 2000; Jordano et al., 2003). This property is related to fundamental ecological limitations: additional interaction partners of a species, for instance frugivores eating the fruits of a specific plant species, will lead to increased competition for eating (i.e. interacting with) that same plant species, disincentivizing additional interactions with it (Montoya et al., 2006). The fact that hub species are nonetheless present, albeit with smaller neighborhoods than in the scale-free model, may be explained by species abundance as for instance a more common prey species may have more predators specializing on it (Montoya et al., 2006).

The right-tail heavy degree distribution of ecological networks appears to play a role in the remarkable robustness to perturbation observed in ecological networks. For instance, the random removal of species typically does not strongly affect the average path length in these networks, indicating their resilience and robustness to random perturbations. This implies reduced second-order extinctions, i.e. species becoming extinct due to the removal of a partner species, to which random networks are noticeably less robust (Montoya and Solé, 2002; Solé and Montoya, 2001). Similarly, complex ecological models can often be parameterized to yield a single stable equilibrium state (McCann, 2000), while many complex differential equation models across domains do not have that property (May, 2001). However, in contrast to random removal, the targeted removal of highly connected species tends to have a high impact on ecological networks and can effectively disrupt their structure (Pimm, 1982). Studies also suggest that indirect interactions, self effects and non-random patterns of weak and strong interactions can be important promoters of stability (Barabás et al., 2017; Kokkoris et al., 2002; Menge, 1995).

Certain network motifs (see Figure 1.6 B), for instance tri-trophic food chains and omnivory, have furthermore been found over- or underrepresented in many ecological networks (Bascompte, 2009), and theoretical evidence suggests that the prevalence of network motifs, as well as the higher-order composition of networks with different motif types, may affect network stability (May, 1972). Architectural properties of networks like these may indeed influence the biodiversity that can be supported by an ecosystem (Bastolla et al., 2009).

Ecological networks tend to be hierarchical and nested, which implies that specialist species typically interact with a subset of the interaction partners of the next more

[10]However, the widely-reported scale-free property of various types of biological networks has been challenged on statistical terms (Lima-Mendez and Helden, 2009)

general species (Bascompte et al., 2003; Clauset et al., 2008; Woodward et al., 2005). For instance, a generalist predator tends to also prey on all species that more specialized predators target, which in turn target a superset of the next more specialized predators, and so forth. This property results in highly asymmetric and cohesive networks, with a stable core of interactions around which all others are organized (Bascompte et al., 2003). Such architectures may confer improved perturbation responses, for instance by reducing the chance of a species becoming isolated upon removal of other species, and thus increase the survival of in particular rare species (Bascompte et al., 2003). Interestingly, many properties of ecological networks can furthermore be reproduced by only accounting for hierarchy in a comparatively simple model (Clauset et al., 2008).

Ecosystems consisting of multiple habitats tend to be compartmentalized, where compartments (or modules) correspond to distinct habitats. There is evidence that mutualistic networks tend to be more nested than compartmentalized, while antagonistic networks have a tendency towards more compartmentalization (Fortuna et al., 2010). The situation is however complicated by a number of confounders, as well as the lack of suitable methods for consistently defining habitats, thus the overall importance and prevalence of compartments remains unclear (Montoya et al., 2006).

Strong phylogenetic signals exist in ecological networks with evidence suggesting more pronounced patterns within compartmentalized than in nested topologies, and in consequence also increased phylogenetic effects in antagonistic compared to mutualistic networks (Rohr and Bascompte, 2014). Furthermore, higher phylogenetic constraints seem to be placed on prey compared to predators in antagonistic networks, and on animals compared to plants in mutualistic networks (Rohr and Bascompte, 2014). These pronounced phylogenetic signals within ecological networks are in line with the proposal that species interactions strongly affected and shaped the expansion of biodiversity, with considerable impact on coevolution (Thompson, 1994; Thompson, 2005). Indeed, indirect and asymmetric species interactions appear to have a profound catalyzing effect on coevolution, for instance by facilitating long-term coexistence (Bascompte et al., 2006; Guimarães Jr et al., 2017).

Another important factor in ecosystems are keystone species, i.e. taxa with a critical effect on ecosystem function relative to their abundance (Power et al., 1996), a concept first introduced in 1966 by Robert T. Paine in the context of the predatory starfish *Pisaster ochraceus* (Paine, 1966). While commonly regarded as species with positive ecosystem effects, wider keystone definitions also incorporate species that fundamentally alter their ecosystem through a disruptive negative impact. Classic examples of positive keystone species include apex predators, such as wolves (*Canis lupus*), because the removal of these predators—despite their comparatively small numbers—has a transformative effect on ecosystems (Power et al., 1996). This impact is mainly mediated by prey being left unchecked and free to multiply, with devastating

effects on other wildlife. The exceptional importance makes keystone species prime targets for land use, conservation and biodiversity restoration efforts (Mills et al., 1993). In ecological networks, keystone species are typically central nodes with many interactions, or nodes occupying key positions whose removal would disrupt the network (Jordán, 2009).

While historically, community assembly is expected to be guided by species-specific traits, niche-adaptations and species-species interactions, Stephen Hubbell's "unified neutral theory of biodiversity and biogeography" (Hubbell, 2001) (or "neutral theory of ecology" in short, not to be confused with the "neutral theory of molecular evolution" by Motoo Kimura, (Kimura, 1983)) has been proposed as a controversial alternative model. It assumes that species differences and species-species interactions are inconsequential for the reproductive success of individuals and instead postulates that simple birth and death patterns, in conjunction with migration from a regional metapopulation, can alone explain observed local species abundances. While some communities show a good fit with this model (Hubbell, 2001; Woodcock et al., 2007), the overall evidence suggests that neutrally assembled communities are rather rare (McGill et al., 2006; Wootton, 2005). However, the neutral theory has nonetheless been established as a valuable null model to test the relative importance of niche-adaptation and species-species interactions (Gotelli, 2000; Leigh, 2007; Rosindell et al., 2011).

As explained in subsection 1.1.1, microorganisms also form complex interaction networks, driven by various mutualist and antagonist interaction types and in many cases dwarfing ecological systems of plants and animals in terms of size and complexity. However, the study of both general and specific properties of these systems is still in its infancy (Röttjers and Faust, 2018), and the main underlying reason is the difficulty of generating network models of these systems in the first place. While identifying all pairwise species interactions is already a daunting task for plant-animals networks (Morales-Castilla et al., 2015), the drastically increased species diversity (Locey and Lennon, 2016; Schloss et al., 2016), as well as complex environmental requirements (Overmann et al., 2017), make this problem even more formidable for microbial ecosystems. Further complications include the spatial and temporal scales that microbes operate on, as well as their morphological homogeneity and the notorious difficulty of distinguishing microbial species (see subsection 1.1.3), all of which necessitate specialized experimental methods and/or theoretical considerations. Statistical associations and mathematical modeling techniques, which can utilize the wealth of currently available metagenomic and genomic data, may however be used to complement initial experimental efforts and allow first glimpses into the structure and functioning of microbial ecosystems (Faust and Raes, 2012; Layeghifard et al., 2017; Röttjers and Faust, 2018).

Networks predicted by such statistical approaches typically share important

properties with ecological networks of plants and animals, such as right-tail heavy degree distributions and the small-world property (Faust and Raes, 2012). Many studies on such inferred network models have furthermore focused on identifying microbial keystone taxa, for instance in human body sites, soil communities and marine environments (Banerjee et al., 2018). Just as their macroscopic counterparts, these species exert their transformative impact on microbial communities independently of their abundance, in contrast to dominant taxa that affect ecosystems solely through their abundance (Banerjee et al., 2018). Many studies in pursuit of microbial keystone taxa so far narrowly focussed on highly connected hubs within co-occurrence networks (see Figure 1.6 B), whose predictivity of keystone status is unclear (Berry and Widder, 2014; Freilich et al., 2018). Nonetheless, results were in some cases replicated in multiple contexts and with different methods (Banerjee et al., 2018). Furthermore, empirical evidence exists for the keystone status for certain microbial groups, for instance several keystone-pathogens (see the "keystone-pathogen hypothesis", Hajishengallis et al. (2012)) involved in oral inflammation, disease-related colonization of the human gut and plant root dysbiosis (Agler et al., 2016; Curtis et al., 2014; Hajishengallis et al., 2011). Other experimentally verified keystone taxa include nitrogen-fixing bacteria in plant root nodules (Heijden et al., 2006).

Considerable research has been conducted on the stability and resilience of microbial ecosystems (Moya and Ferrer, 2016; Shade et al., 2012). In longitudinal studies, gut communities showed pronounced long-term stability and the tendency to relapse into their initial structure after perturbation (for instance antibiotic or dietary interventions), furthermore indicating marked resilience (Claesson et al., 2011; David et al., 2014a; Dethlefsen and Relman, 2011; Zoetendal et al., 1998). Nonetheless, also permanent and sometimes drastic shifts in community composition have been observed following perturbation (Allison and Martiny, 2008; Hernández et al., 2013; McFarland, 2008). Studies furthermore found community function to be highly conserved in various ecosystems, for instance marine, human- and plant-associated environments, despite pronounced species differences between habitats (Huttenhower et al., 2012; Louca et al., 2018; Moya and Ferrer, 2016; Sunagawa et al., 2015; Turnbaugh et al., 2009). This suggests that pronounced functional redundancy across taxa may be a widespread property in microbial communities, with important implications for ecosystem resilience.

Recent results furthermore indicate that community assembly may follow repeatable patterns of emergence and self-assembly on a coarse-grained taxonomic scale (Goldford et al., 2018). These patterns appear to be driven and stabilized by higher-order cross-feeding interactions between microbes, suggesting the presence of stable equilibria in microbial ecosystems. In contrast, several theoretical and data-driven models suggest that widespread competition may be necessary for stability in microbial

ecosystems in the murine gut (Coyte et al., 2015; Stein et al., 2013). Evidence for multi-stability, i.e. the existence of multiple stable community states within the same habitat, is furthermore discussed for the human gut microbiome (termed "enterotypes"; Arumugam et al., 2011; Gonze et al., 2017; Knights et al., 2014).

A repeatedly observed phylogenetic pattern in microbial co-occurrence analyses is phylogenetic assortativity, i.e. the increased probability of associations between phylogenetically more closely related groups (also see Figure 1.6 B, albeit for a different type of assortativity), which was replicated across a variety of data sets and methods (Chaffron et al., 2010; Faust et al., 2012; Kurtz et al., 2015). While widespread competition was suggested as a possible explanation (based on additional genomic evidence, Chaffron et al. (2010)), the lack of experimentally verified interactions has so far precluded robust validation of this pattern.

Hubbell's neutral theory has also been applied to microbial communities, with good fits in some studies (Woodcock et al., 2007) and mixed results[11] or negligible explanatory power (Cregger et al., 2018) in others. Both neutral and non-neutral processes may thereby contribute to microbial community assembly (Faust and Raes, 2012).

Strong modularization is also often observed in co-occurrence networks, which was attributed to niches or abiotic factors in some studies, but the nature of such compartments is not always clear (Röttjers and Faust, 2018). Similarly, the prevalence and importance of network motifs (see Figure 1.6 B)—studied to some extent in animal and plant networks (Bascompte, 2009)—remains underexplored in microbial interaction networks, and the potential biological meaning of previously reported motifs remains to be elucidated (Röttjers and Faust, 2018).

1.1.3 Microbial systematics and the species problem

> *Imagine walking out in the countryside and not being able to tell a snake from a cow from a mouse from a blade of grass, that's been the level of our ignorance.*
>
> —Carl Woese

Describing and categorizing the enormous diversity of microbial life is a gargantuan task. Through recent improvements in molecular sequencing approaches (see subsection 1.2.2), the true extent of undescribed microbial biodiversity became dramatically apparent. While in 1987 only 12 phyla were known (Madigan et al., 2014), 2017 already saw a total of 80 named bacterial and 26 named archaeal phyla, out of which only 30 and 3, respectively, had cultured representatives

[11] for instance applying only to generalist but not specialist species or being restricted to one environment (Liao et al., 2016; Sloan et al., 2006)

(Overmann et al., 2017). Furthermore, the majority of genes identified through culture-independent sequencing approaches, in particular in environmental samples, are currently undescribed (Overmann et al., 2017). While numbers of described species steadily increase (currently at around 12'000 characterized species), new additions mostly fall within higher taxonomic groups that already have cultured representatives, and thereby only increase the depth but not the breadth of available cultures (Overmann et al., 2017). These observations clearly highlight the importance of time-consuming yet unavoidable isolation, cultivation and description work to systematically approach this staggering wealth of diversity. While culturing and description still plays a catch-up game, and the gap indeed seems to be increasing (Overmann et al., 2017), high-throughput culturing approaches may help address this problem (see subsection 1.2.1).

Reasons for this gap include the high burdens for naming a new species: tellingly, the overall costs of validly describing a single bacterial isolate have been estimated at 9'836 € (Overmann, 2015). To validly assign a binomial Linnean name to a newly discovered species, a number of prerequisites must be met (Madigan et al., 2014). Firstly, a detailed description of characteristics and, importantly, distinguishing traits compared to other species, must be published in an adequate journal (traditionally the Journal of Systematic and Evolutionary Microbiology (IJSEM)). Secondly, viable pure cultures of the new organism (i.e. type strains) must be sent to at least two culture collections in different countries. The new species will then be included in central taxonomic reference resources, for instance the "List of Prokaryotic names with Standing in Nomenclature", completing the naming process.

However, through the advent of molecular methods, additional sequence-based information is nowadays also used (and required) for species delineation. DNA-DNA hybridization experiments, in which the nucleotide similarity between two genomes is measured experimentally, showed that average nucleotide identity (ANI) scores of more than 70% (given standardized conditions) corresponded well to species described at the time (Wayne et al., 1987). However, shortcomings of DNA-DNA hybridization include several technical difficulties, for instance that results can change depending on which genome is used as the probe and which as the target, that the method is intransitive (genome A can have ANI >=70% to genome B and genome C, but genome B can have ANI <70% to genome C) and that the 70% cutoff was based on pre-existing species assignments that lacked a theoretical foundation (Achtman and Wagner, 2008). Nonetheless, this method provided a much-needed molecular basis for distinguishing species as opposed to depending on less reliable (and often hard to measure) phenotypical traits.

Because DNA-DNA hybridization is expensive and laborious, alternative thresholds based on nucleotide identity of the 16S rRNA gene were proposed and became the de-facto standard for approximate species assignment. Already proposed and used

by Woese and colleagues in the earliest days of molecular phylogenetics (Fox et al., 1977; Woese and Fox, 1977), the 16S gene has a number of beneficial traits for taxonomic classification (Hugerth and Andersson, 2017; Madigan et al., 2014): (i) it is ubiquitously found across all known organisms, (ii) it features both highly conserved and highly variable ("hypervariable") regions, thus conserved regions can be used as reliable targets for primer binding, while hypervariable regions provide enough variation to inform taxonomic distinction, (iii) it is rarely transferred horizontally and (iv) it can easily be handled experimentally due to its small size, which was an advantage especially in the early Woesian days.

Comparison of 16S rRNA identity with DNA-DNA hybridization experiments showed a good correspondence between the two, with a threshold of 97% typically corresponding to a hybridization ANI of 70%, leading to the suggestion of the 97% 16S identity threshold to classify organisms as distinct species. Conversely, a 16S rRNA identity above 97% would be an indication for both organisms coming from the same species, albeit this required further confirmation through whole-genome hybridization because some distinct species showed 16S rRNA identity above 97%. A more recent re-evaluation of the 97% threshold, based on more recent experimental data that included more species, however found that the threshold for species distinction should be more appropriately set at 98.7% (Stackebrandt, 2006). These thresholds are furthermore used to assign sequences, for which no additional information is available and which do not match any described reference species, into sequence-based phylotypes (for instance operational taxonomic units (OTUs), see subsection 1.2.3).

The 16S rRNA gene is additionally sometimes used to assign a provisional "Candidatus" name to uncultured microbes with largely unknown phenotypes. Given sufficient 16S rRNA difference to known species, as well as a basic description of species traits as far as available (e.g. based on genomic information) and an *in situ* hybridization protocol for specific detection, such preliminary names can be validly assigned (Murray and Stackebrandt, 1995) and are becoming increasingly popular due to the mass of novel groups and the painstaking work that can be required for culturing (see subsection 1.2.1). Similar to 16S rRNA, the internal transcribed spacer (ITS) region, located between the 16S rRNA and 23S rRNA genes in archaea and bacteria and between the 18S and the 5.8S rRNA genes in eukaryotes (ITS1), has been found to improve distinction between certain taxa, for instance in fungi and plants (Baldwin et al., 1995; Peay et al., 2008). It has furthermore been used to increase resolution in particular bacterial groups, such as cyanobacteria and SAR11 (Sullivan et al., 2003; Zhao et al., 2013).

Through more advanced sequencing methods, which provide information on multiple genes at once, it became clear that the 16S rRNA gene or ITS region alone can be limited in their taxonomic resolution. For instance, distinct *Bacillus* species with

different pathogenicity patterns and ecology have nearly identical 16S rRNA genes (Sacchi et al., 2002) and thus require particular care for proper species delineation. But also within the same species, different strains with almost identical 16S rRNA sequence can differ drastically: for instance, the K-12 strain of *E. coli* is generally harmless, while the O157:H7 strain can cause fatal infections (Perna et al., 2001). Some organisms furthermore carry more than one 16S rRNA gene, which additionally complicates comparison and species distinction (Sun et al., 2013; Větrovskỳ and Baldrian, 2013).

To improve upon this situation, information from additional genes can be used, in particular protein-coding housekeeping genes may provide better resolution since they tend to evolve faster. An approach that combines a number of such genes (typically 6-8) is multilocus sequence typing (MLST), which is used especially in medical contexts where rapid identification of pathogenic strains is essential (Madigan et al., 2014). With the availability of more and more microbial genomes, genome-wide comparisons of orthologous genes via gANI (genomic ANI) is also increasingly used to distinguish species. A gANI threshold of 95% was found to best correspond to 70% DNA-DNA hybridization ANI by an initial analysis (Goris et al., 2007), while a more recent study based on more than 10'000 genomes found a threshold of 96.5% to be most compatible with previous hybridization-based species assignments (Varghese et al., 2015).

The advantage of genomic data also became apparent in the context of phylogenetic reconstruction, which aims at elucidating taxonomic affiliations deeper than species level. For instance, an important breakthrough in our understanding of early evolution came through the use of large numbers of protein-coding housekeeping genes for phylogenetic reconstruction. While the original three-domain tree of life, as presented by Woese and colleagues based on 16S rRNA gene information (Woese et al., 1990), featured *Bacteria*, *Archaea* and *Eukaryota* are separate branches, with *Archaea* and *Eukaryota* being sister clades[12], modern revisions of the tree of life draw a different picture. Based on substantially increased numbers and diversity of genomes, as well as information from a large number of conserved housekeeping genes, these trees show *Archaea* to be paraphyletic, with *Eukaryota* evolving from within *Archaea* rather than being a sister group (Cox et al., 2008; Guy and Ettema, 2011). The implications are profound: according to this model, higher organisms don't only share an undefined ancestor with modern *Archaea*, but may have evolved from bonafide archaeal cells.

The previously discussed thresholds, despite their usefulness in operationally describing microbial diversity and providing an essential common language for scientific discourse (Godfray, 2002), are yet merely working definitions that lack a theoretical evolutionary underpinning. To improve on this, ample discussion has focussed on

[12]Through this finding, it furthermore became evident that the term "prokaryotes" cannot be evolutionarily justified since organisms without a nucleus do not form a monophyletic group (Pace, 2006) and this insight remains valid to this day.

the development of evolutionarily sound and consistent species concepts (Lawrence and Retchless, 2009). While a sound species concept, underpinning a consistent species definition, could allow a more accurate description of biodiversity and improve inference of shared traits and evolutionary trajectories, desirable properties of a unifying concept are hotly debated (Achtman and Wagner, 2008; Doolittle and Papke, 2006; Rosselló-Mora and Amann, 2001). Though concepts based on reproductive isolation, such as the classic "biological species concept" proposed by Ernst Mayr (Mayr, 1942), describe the majority of eukaryotic species well, they do not easily extend to bacteria and archaea due to abounding lateral gene transfer (LGT) within (and between) these groups, conveyed through the mechanisms of transduction, conjugation or transformation (Soucy et al., 2015).

The true extent of LGT started being appreciated only more recently through the advent of massive environmental sequencing of whole genomes, in tandem with improved bioinformatic methods for LGT detection (Soucy et al., 2015). Such transfer events can cross species boundaries and even bridge higher-level taxa (see Figure 1.7 A), in many cases leading to drastic switches in phenotype, ecology and consequently evolutionary pressures. This effect blurs species boundaries and makes the archaeal and bacterial tree of life more reticular, in fact putting the concept of a hierarchically organized taxonomy in these groups into question altogether (the "species problem"; Doolittle and Papke, 2006; Ereshefsky, 2010; Hey, 2001). As a result, phylogenies constructed for different genes tell different stories, depending on whether and where such LGT events took place, and evolutionary units defined on these grounds are thereby arbitrary (Doolittle and Bapteste, 2007).

Nonetheless, various concepts have been proposed in hope of addressing this conundrum, for instance based on phenotypic traits, phylogenetic characteristics (such as monophyly), the frequency of lateral gene transfer, ecotypes, metapopulation lineages and others (Achtman and Wagner (2008); see Figure 1.7 C). It is now widely accepted that strong cohesive forces affect many microbial groups, and that talking about "species" of microbes in the context of these groups may be appropriate, while other groups devoid of this property cannot be adequately described as species (Achtman and Wagner, 2008; Doolittle and Zhaxybayeva, 2009; Lawrence and Retchless, 2009; Shapiro et al., 2016). These forces include for instance shared selective pressures, i.e. similar ecological requirements across group members, and intrinsic mechanisms, such as homologous recombination (Shapiro and Polz (2014); see Figure 1.7 B). The latter favors genetic exchange between closely related organisms and thus represents a process that homogenizes groups within species boundaries, separating them from more distantly related groups (Lawrence and Retchless, 2009). Nonetheless, methodological challenges must still be overcome to convert theoretical concepts of cohesive forces into workable species definitions. For instance, consistent

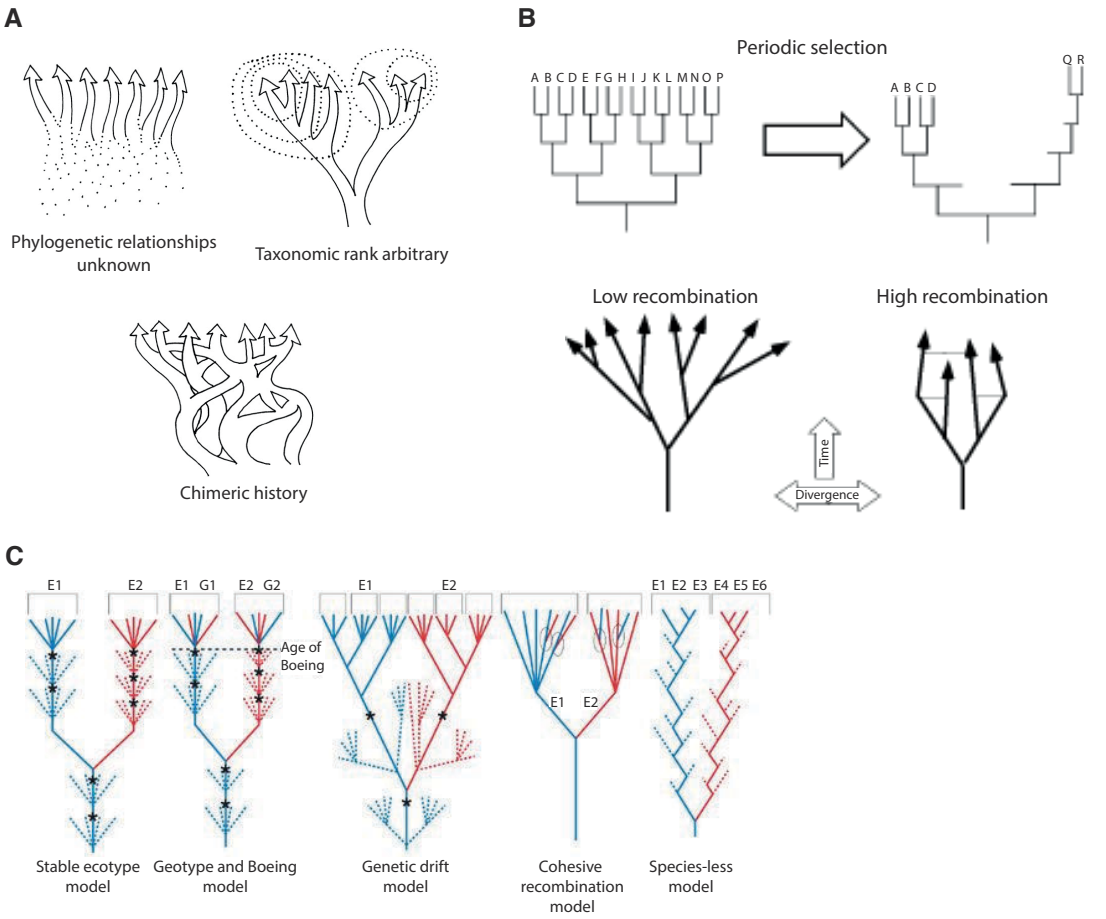


Figure 1.7: Challenges and concepts for species delineation in microbiology. **A** Taxon delineation is generally difficult since the choice of distinguishing traits and thresholds is inherently subjective. For microbial taxa, widespread lateral gene transfer can additionally reduce the resolution of evolutionary inference and blur taxon boundaries. Adapted with permission from Doolittle and Zhaxybayeva (2009). **B** The strongest forces for evolutionary coherence of microbial groups are periodic selective sweeps and homologous recombination, the latter being particularly pronounced between closely related groups and able to create discrete lineages. Adapted with permission from Lawrence and Retchless (2009). **C** An overview of different ecotype models suggested as underpinnings for microbial species definitions. Adapted with permission from Achtman and Wagner (2008), where further details and explanations can be found.

approaches to meaningfully distinguish ecological niches and predict evolutionary trajectories, which may furthermore depend on the group of microbes under study, are not yet available (Achtman and Wagner, 2008; Lawrence and Retchless, 2009).

Apart from accurately defining and classifying microbial species, accurate and consistent systematics of higher-order taxonomic groups are also crucial for many scientific questions. While names are assigned to these groups according to the International Code of Nomenclature of Prokaryotes (Parker et al., 2015), theoretical concepts for deeper taxa are still lacking, as no unambiguous biological mechanisms to maintain group coherence are evident at these higher levels (Lawrence and Retchless, 2009). Historically, phenotypes have been used for classification of higher-order taxa, which lead to increased numbers of splits within more extensively studied groups, independent of actual evolutionary group divergence. To remedy this, genetic consistency and phylogenetic relatedness between organisms started being used more recently to distinguish higher-order taxonomic groups through a data-driven

working definition—in particular, through evidence from phylogenetic trees based on concatenated housekeeping proteins, sometimes complemented by the estimated overlap of orthologous gene content (Brown et al., 2015; Guy and Ettema, 2011; Rinke et al., 2013; Zaremba-Niedzwiedzka et al., 2017).

These sequence-based methods, applied to the staggering amounts of novel genomes from undescribed organisms, made the ambiguity of traditional higher-order groups strikingly apparent. For instance, the highly studied family *Enterobacteriaceae* shows similar evolutionary divergence as many genera (for instance *Bacillus*), while taxa in understudied groups (for instance the family *Synergistaceae*) show substantially higher divergence than more deeply studied taxa of the same rank (Parks et al., 2018). Lately, studies have been conducted towards creating a more consistent, sequence-based taxonomy that encompasses both cultured and uncultured organisms, for instance based on 16S rRNA (Yarza et al., 2014) or, more robustly, on phylogenies inferred from concatenated proteins (Parks et al., 2018).

The latter effort by Parks et al was based on 21'943 dereplicated bacterial genomes, 40% of which constituted metagenome-assembled genomes (MAGs) and single-cell genomes from uncultured organisms, leading to a drastic increase of the diversity informing their taxonomy. By also applying an approach for robust calibration of evolutionary divergence between inferred groups, the authors equalized evolutionary divergence rates within ranks and thus provided a major step towards a more consistent genome-based taxonomy (Garrity, 2016; Hugenholtz et al., 2016), compiled into the Genome Taxonomy Database (GTDB, Parks et al. (2018)).

This taxonomy led to the re-assignment of taxonomic groups for 58% of all genomes, including many deeply-branching changes. Striking examples include the reorganization of the genera *Clostridium* and *Bacillus* into 121/29 and 81/25 genera and families, respectively, as well as the reclassification of the class *Betaproteobacteria* as an order within the class *Gammaproteobacteria*, and reclassification of the candidate phyla radiation (CPR, including 65 proposed candidate phyla) into a single phylum. Albeit the genomes used for this approach still only capture a fraction of the expected microbial diversity, and revisions based on new data are thereby expected, such data-driven approaches provide a consistent way forward to tame the juggernaut of microbial diversity.

1.2 Methods for studying microbial ecology

The world is full of magic things, patiently waiting for our senses to grow sharper.

—William Butler Yeats

1.2.1 Culture techniques

The first observation of a microorganism was reported by Robert Hooke already in 1664, followed by the discovery of bacteria by Antoni van Leeuwenhoek (1676), enabled through his fundamental improvements of early microscopes (Madigan et al., 2014). Much research at the time was conducted via observation of natural samples under light microscopes, until the first microbial culturing techniques were developed in the middle to late 18th century, pioneered by Robert Koch and Louis Pasteur (among others) and motivated by clinical applications (Madigan et al., 2014). Soon after, Sergei Winogradsky and Martinus Beijerinck championed culturing techniques for the study of environmental microbes from soils and aquatic habitats, which lead to a number of breakthroughs, such as the first description and culture of nitrogen-fixing bacteria. This milestone of environmental microbiology was enabled through innovative enrichment culture devices, such as the Winogradsky column (Madigan et al., 2014). Beijerinck furthermore created the first pure (or axenic) cultures of various terrestrial and aquatic species through enrichment culture, most prominently *Azotobacter chroococcum* via a selective N_2 medium. This concept of growing environmental microbes in a medium that is selective for the target species and counter-selective for others remains the standard strategy for enrichment culture to this day (Madigan et al., 2014).

Ever since, thousands of specialized protocols detailing specific growth media and conditions have been developed, allowing the laboratory culture of tens of thousands of microbial strains from a variety of microbial phyla (Overmann et al., 2017). Nowadays, cultured strains can be ordered from central culture repositories, such as the German Collection of Microorganisms and Cell Cultures (DSMZ[13]), and culture protocols are available within associated databases, like for instance the KOMODO resource provided by the DSMZ. To facilitate finding the right conditions for an uncultured strain, KOMODO can furthermore suggest media and conditions based on phylogenetic relationships to cultured strains within a recommender system, fed with the wealth of available culturing data (Oberhardt et al., 2015). Protocols include for instance nutrient composition (rich vs. poor), oxygen content, salinity, pH and ionic concentrations, as well as temperature and incubation times, with the aim of providing conditions closely

[13]Deutsche Sammlung von Mikroorganismen und Zellkulturen

mimicking the natural environment (*in situ* cultivation, Kaeberlein et al. (2002)). A remarkable success of *in situ* cultivation is the isolation and pure culture of the first obligate piezophilic and hyperthermophilic archaeon (*Pyrococcus* CH1). This strain was isolated from deep-sea hydrothermal vent samples at 4100 m depth and subsequently grown under *in situ* conditions of 42 MPa pressure and 95 °C temperature within specialized bioreactors (Zeng et al., 2009).

Species of interest can be isolated from samples (possibly following enrichment culture) by either simple techniques, such as the streak plate method (see Figure 1.8 A), which is restricted to species growing on agar, or more advanced approaches, such as the Laser tweezer and Flow cytometry-based methods (Madigan et al., 2014). The latter sort cells according to morphological characteristics or molecular fluorescent markers and thus allow for the more specific isolation of target organisms. In all cases, the purity of a culture may subsequently be verified through growth on other media (selective for contaminants) and microscopic observation of community characteristics (e.g. morphology or stainings patterns). This coarse method can, however, be unreliable due to the morphological similarity of even distantly related species. In addition, stochastic expression patterns may lead to phenotypic heterogeneity even in clonal populations, necessitating more specific methods (Madigan et al., 2014). One such approach are fluorescence *in situ* hybridization (FISH) probes that target the 16S component of the small 30S ribosomal RNA subunit (or 16S rRNA in short, see subsection 1.1.3) and can thus differentiate phenotypically similar species.

Other typical requirements include the distinction of live and dead cells, which can be achieved through viability staining with special dyes that penetrate cells only when membrane integrity is lost, and the monitoring of specifically engineered strains within a natural community, addressable by introducing genes for autofluorescence (for instance with green fluorescent protein (GFP) reporters) into the strain of interest (Madigan et al., 2014). Specific subpopulations in mixed communities can further be followed with modern fluorescence methods (e.g. FISH), which additionally enable multiparametric analysis of such communities by measuring many variables of interest in parallel (Madigan et al., 2014).

Similarly, the metabolic activity of both cultured and natural communities can be studied through a range of methods, for instance by community-wide approaches like chemical assays, radioisotope methods, electrochemical (as well as more recently "living") microsensors and isotope fractionation. Also more modern methods with single-cell resolution, which couple activity to phylogenetic diversity, are now available. These include nanoscale secondary ion mass spectrometry (NanoSIMS), microautoradiography-FISH (MAR-FISH) and stable isotope probing (SIP), and allow for dedicated studies on the metabolic potential of single strains within a community of interest (Madigan et al., 2014).

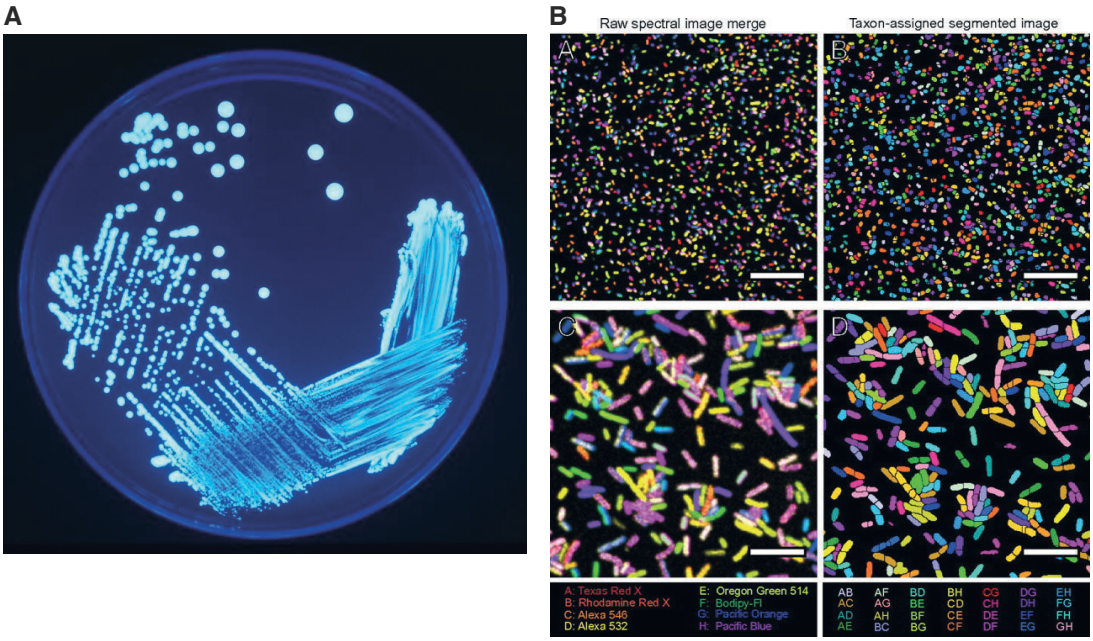


Figure 1.8: Examples of culture and monitoring techniques. **A** The streak plate method is a classic technique for isolating and culturing microbes. Credit due to Wikimedia Commons (2005). **B** CLASI-FISH is an advanced monitoring approach for combinatorial fluorescent labeling and the simultaneous tracking of many distinct microbial groups within complex communities. Adapted with permission from Valm et al. (2011).

Pure cultures serve multiple purposes. First and foremost, they are an invaluable tool for the detailed study of specific organisms under controlled laboratory conditions, which can be suitably altered depending on the scientific question. In addition, pure cultures deposited in a central culture collection are currently obligatory for validly describing a species with a binomial Linnaean name (Lapage et al., 1992). Such strain cultures are furthermore needed to fulfill Koch's postulates and thus identify causative agents of infectious diseases, as well as for the detailed study of pathogenicity mechanisms (Overmann et al., 2017). Cultured strains with well-annotated genomes are also an important resource for functional predictions in metagenomic and single-cell genomics analysis since novel and complex traits (for instance autotrophic growth on unusual compounds) or atypical enzyme kinetics typically cannot be inferred from genomic information alone (Overmann et al., 2017). Indeed, culturing coupled with biochemical analysis has disproved genomic predictions in several cases (Overmann et al., 2017).

For consistent coverage of microbial diversity, both breadth and depth are important in cultivation: ideally, one wants to get at least one member for each group (breadth), but in addition also multiple members per group to characterize it more fully (depth) (Madigan et al., 2014). A profound obstacle to cultivation efforts, in particular concerning the breadth of coverage, is what has become known as the "great plate count anomaly" (Staley and Konopka, 1985). This term describes the concerning mismatch between the diversity observable in natural samples under the microscope vs. the diversity present in laboratory cultures derived from the same samples. In

many cases, species that grow well in cultures on artificial media are only present in low numbers within the original samples, and are thus likely of negligible ecological relevance, while other more relevant species do not grow at all (Madigan et al., 2014). This phenomenon induces a profound enrichment bias and in consequence, knowledge is at present largely missing about the vast majority of known phylotypes (i.e. known high-identity sequence clusters of microbial marker genes (e.g. the 16S rRNA gene), see subsections 1.1.3 and 1.2.3).

Reasons for this bias are manifold. For instance, microbes growing fast only in culture ("weed" species) may be less fastidious than other community members, for example with regards to physicochemical conditions or microbial interactions partners. Furthermore, many weed species are copiotrophs (i.e. prefer nutrient-rich conditions) that disproportionately benefit from artificially nutrient-rich media and thus outcompete oligorophic microbes (i.e. which prefer nutrient-poor conditions). Indeed, oligotrophic organisms are so adapted to nutrient-poor conditions, as typical for many habitats (e.g. soils and marine environments), that they are in fact inhibited by rich growth media. Consequently, most cultivated marine microbes are copiotrophs, despite nutrients being scarce in marine environments, indicating a concerningly poor fit between culturing media and natural conditions (Overmann et al., 2017). The effect of media-dependent competitive advantages can be clearly observed in dilution cultures, where successive dilutions of an inoculum yield different species, induced by the removal of weed species during the dilution step (Madigan et al., 2014).

Other problematic factors include the slow growth rates of many uncultured microbes, which make trial and error approaches in cultivation experiments particularly time intense. Another challenge is posed by the small size of certain uncultured groups, such as the ultramicrobacteria, which necessitates filters with very small pore sizes for isolation. Use of such filters for instance lead to the first isolation and pure culture of a representative from the phylum *Endomicrobia* (Geissinger et al., 2009). Stochastic fluctuations of gene expression may pose another obstacle: even clonal laboratory populations can display phenotypic heterogeneity, where only a subpopulation expresses the genes or pathways essential for growth in a given culturing setup, a phenomenon caused by transcriptional stochasticity (Ackermann, 2015). If gene activity profiles do not match the culture conditions, strains may thus arbitrarily fail to grow. Other problems include specific adherence characteristics of some uncultured species, which thus require particular surfaces for successful cultivation (Gich et al., 2012), and the temporal variability of nutrient concentrations in natural habitats, which need to be reproduced for certain species to grow (Overmann, 2005).

Since enrichment bias has been recognized more widely, an array of methods has been developed to cultivate groups that are more representative of natural variety. As previously mentioned, dilution culture can for instance help separate

target organisms from highly competitive weed species (Madigan et al., 2014). More recent developments include special culture chips that detect bacteria even at low concentrations (Bouguelia et al., 2013) and traps that specifically capture particular hard-to-culture groups, such as filamentous bacteria (Gavrish et al., 2008).

Emphasis in modern methods is often placed on scale: high-throughput culturing approaches now allow for massively parallel cultivation experiments, including thousands or millions of combinations of individual species, media and physicochemical conditions (Overmann et al., 2017). Cells from an inoculum can for instance be sorted into separate wells with different media, followed by incubation and parallel screening for groups of interest (possibly aided by fluorescence or sequencing approaches, see earlier paragraphs and subsection 1.2.2; Madigan et al. (2014)). This provides distinct advantages for culturing slow-growing species, with doubling times of several months or more, as many different conditions can be tested in parallel (Madigan et al., 2014). A high-throughput dilution culture approach for instance allowed the first pure culture of the ubiquitous, but notoriously fastidious, marine clade SAR11 (Rappé et al., 2002). Other high-throughput methods include the isolation chip (ichip, Nichols et al. (2010)), optimized for highly parallel *in situ* culture, the micro-Petri dish with up to a million miniature wells for separate cultivation (Ingham et al., 2007), as well as so-called culturomics approaches specialized on isolation and culture of novel microbial groups in human and mouse guts (Lagier et al., 2018).

Another class of culture methods are efficient microfluidic devices, such as a cultivation platform with highly sensitive tuning of conditions and flows for single cells (Grünberger et al., 2015), a microdispenser system optimized for the culture of planktonic bacteria at scale (Bruns et al., 2003) and the Microbiome-on-a-Chip for the culture of complex communities, which additionally facilitates study of their impact on host systems (Stanley and Heijden, 2017). To reliably identify (possibly rare) target organisms in these high-throughput systems, matrix-assisted laser desorption with time-of-flight (MALDI-TOF) approaches can be used to detect known species (Seng et al., 2013), while combinatorial barcoding of 16S rRNA amplicons also allows the identification of unknown phylotypes (Camarinha-Silva et al., 2014).

Other breakthroughs came from methods that allow the prediction of necessary culture conditions. For instance, omics approaches (in particular (meta-)genomics, -transcriptomics and -proteomics) may help estimate the metabolic potential and requirements of target species and thus inform the choice of culture media and conditions. As an example, genomic analyses predicted specific metabolic requirements and a dependency on low-oxygen for the intracellular pathogen *Coxiella*, which informed its first successful cultivation under laboratory conditions (Omsland et al., 2009). Similarly, genomic predictions identified the requirement for reduced environmental sulfur by SAR11 (Tripp et al., 2008). Modern tools for activity

measurement (NanoSIMS, MAR-FISH, SIP) may also be used to identify metabolic requirements, for instance by tracing, which substrates from the medium are incorporated into target cells (Eichorst et al., 2015; Könneke et al., 2005; Madigan et al., 2014). Chemotaxis assays are an alternative for motile species, which tend to accumulate in the proximity of preferred compounds within the experimental setup (Overmann, 2005). Novel phylogenetic groups may also be directly targeted through ecologically-motivated statistical methods, which computationally identify promising biotic and abiotic factors by associating these factors with target species in culture-independent environmental data sets (Foesel et al., 2014).

Another important aspect considering the improvement of cultivation efforts is the identification of obligate interaction partners of target species. Colonies in cultures obtained directly from environmental samples were found to be significantly more heterogeneous than expected by chance (Kenters et al., 2011), indicating intimate ecological relationships that translate all the way through to culture. In line with this observation, several studies found that many newly cultured microbes could only be grown on artificial media when in co-culture with interaction partners, but not in isolation (Hahn, 2009; Kaeberlein et al., 2002). D'Onofrio et al frequently observed large colonies on Petri dishes that had smaller and phenotypically diverse colonies in close proximity ("satellites"), which lead them to hypothesize a metabolic dependency of the satellite species on the central colony (D' Onofrio et al., 2010). Subsequently, the authors identified the siderophore Enterobactin as the growth factor responsible for these successful co-cultures, and incorporation of this compound into media allowed the pure culture of several novel groups, for instance the *Verrucomicrobia*. Another currently underexplored but promising candidate for growth factors are autoinducers, which play an important role in quorum sensing (Overmann et al., 2017).

Despite these successes, however, targeted optimization of media for pure culture with specific shared compounds often is unsuccessful, in which case co-culture remains necessary (Dedysh, 2011; Hahn, 2009; Kaeberlein et al., 2002; Lage and Bondoso, 2012). An example for this is the cultivation of certain endosymbionts and parasites with complex host dependencies (Fröstl and Overmann, 1998; Pagnier et al., 2015). Methods that allow the spatial separation of partners, while still permitting the flow of exchanged metabolites, may be useful to mitigate some of the practical difficulties that come with obligate co-culture dependencies (Zengler et al., 2002). Importantly, some of the previously mentioned high-throughput techniques can also be used in a co-culture setup to systematically attempt culture for large numbers of species combinations (Overmann et al., 2017), for instance via specifically optimized microfluidic devices (Park et al., 2011).

Apart from enabling culture of fastidious microbes, co-culture also allows the study of ecological interactions under controlled conditions, which was pioneered by George

Francis Gause in the first competition experiments during the 1930s (Gause, 1936). To obtain direct evidence for metabolite flows between interaction partners in such experiments, isotope-labeling methods such as NanoSIMS and MAR-FISH can again be used for high-resolution monitoring of the cells (Madigan et al., 2014).

Culturing methods furthermore allowed the creation of synthetic communities with conveniently reduced complexity. Albeit these are simplified approximations to real microbiota, they can nonetheless be informative for studying the effect of host factors (e.g. genetics) on the microbiome, as was for instance done successfully in plants and the mouse gut (Agler et al., 2016; Bodenhausen et al., 2014; Desai et al., 2016). Synthetic communities have furthermore been used to validate computationally predicted interaction networks, again in the context of plant and mouse gut-associated microbiota (Agler et al., 2016; Faith et al., 2010; Vorholt et al., 2017). If species of interest can be cultured *in vitro*, high-throughput systems, such as the synthetic community system proposed by Chodkowski and Shade, can be used to measure large numbers of microbial interactions via exometabolite profiling (Chodkowski and Shade, 2017). Monitoring species in complex interacting communities is furthermore possible through combinatorial fluorescence labeling approaches, for instance combinatorial labeling and spectral imaging - FISH (CLASI-FISH, Valm et al. (2011)), which allows the simultaneous tracking of hundreds of species and can detect complex, temporally resolved interaction patterns (see Figure 1.8 B). Culturing approaches may furthermore help validate general network features, for instance keystone species, by systematically removing these species from *in situ* cultured communities and measuring the subsequent ecosystem impact.

1.2.2 Sequencing technologies

With the description of the molecular structure of DNA by Watson and Crick (Watson, Crick, et al., 1953), building on crucial results from Rosalind Franklin and co-workers (Klug, 1968), a new era of sequence-based biology was born. Decades later, groundbreaking work lead by Carl Woese resulted in the very first application of an early version of Sanger sequencing to the 16S rRNA of unusual methane-producing microbes, which constituted a pioneering use of sequence information for phylogenetic classification. Through subsequent comparison to bacterial 16S rRNA sequences, it became soon apparent that Woese and coworkers had discovered the third domain of life: the *Archaea* (Woese and Fox, 1977). Around the same time, Frederick Sanger and colleagues further developed their sequencing approach into what is known today as the Sanger sequencing method (Sanger et al., 1977). In combination with the polymerase chain reaction technique (PCR), first firmly established by Kary Mullis in 1983 (Bartlett and Stirling, 2003), this technology quickly became the foundation

for a revolutionary, sequence-based microbiology, enabling investigators to routinely measure microbial diversity directly within samples in a culture-independent fashion (Pace et al., 1986; Stahl et al., 1985; Ward et al., 1990). Tellingly, both methods were awarded Nobel Prizes in Chemistry in 1980 and 1993, respectively. Typical steps of this amplicon-based workflow still prevail to this day and include the extraction of microbial DNA, PCR amplification, cloning and Sanger sequencing.

However, this methodology depended traditionally on time and money intensive cloning and was furthermore reliant on slow, sequential PCR and sequencing. This strongly limited the number of samples and depth per sample, restricting analysis to carefully picked environments and conditions, and leading to the omission of rare taxonomic groups (Madigan et al., 2014). Mitigation of this problem came from yet another sequencing technology revolution, termed "second-generation" or "high-throughput" sequencing, at the beginning of the 21st century. This new class of sequencing approaches allowed the massively parallel amplification and sequencing of microbial samples, cutting the cost and time needed per sequenced read drastically compared to traditional sequential approaches[14]. Modern high-throughput methods allow the sequencing of up to hundreds of billions of sequences within days, for prices below 100 USD per Gbp (giga base pair) (Goodwin et al., 2016). The most prominent technologies from this generation are 454 Pyrosequencing (now discontinued), SOLiD and Illumina, with Illumina MiSeq (typically used for amplicon sequencing) and HiSeq (shorter reads and higher throughput than MiSeq, typically used for metagenomic sequencing) currently dominating the market (Goodwin et al., 2016).

Cheaper alternatives to high-throughput sequencing, albeit less comprehensive and resolved, include the DNA Microarray-based phylochip and GeoChip, which allow the profiling of phylotypes and functional genes, respectively (Madigan et al., 2014). Apart from taxonomic profiling, high-throughput sequencing also enabled routine assemblies of new genomes, which unveiled a surprising amount of genomic plasticity within genomes of the same species and lead for instance to the insight that pan-genomes (the full gene repertoire utilized by any member of a species) dwarf the core genomes (genes present in all genomes of a species) for many species (Medini et al., 2008).

A problem for taxonomic profiling via amplicon sequencing, which affects both Sanger sequencing and high-throughput sequencing, are widely recognized biases introduced by the PCR amplification step (Klindworth et al., 2013; Tremblay et al., 2015). These systematic errors can lead to overestimation of the relative abundance of correctly amplified groups while underestimating relative abundances of (or even missing) groups with suboptimal matches. A concerning example are *Nanoarchaeota* and certain nitrifiers, which are not properly amplified by popular primers (Diwan et al., 2018; Huber et al., 2002). Even supposedly universal primers, designed to capture

[14] costs per base dropped 10'000 fold between 2001 and 2011 (Madigan et al., 2014)

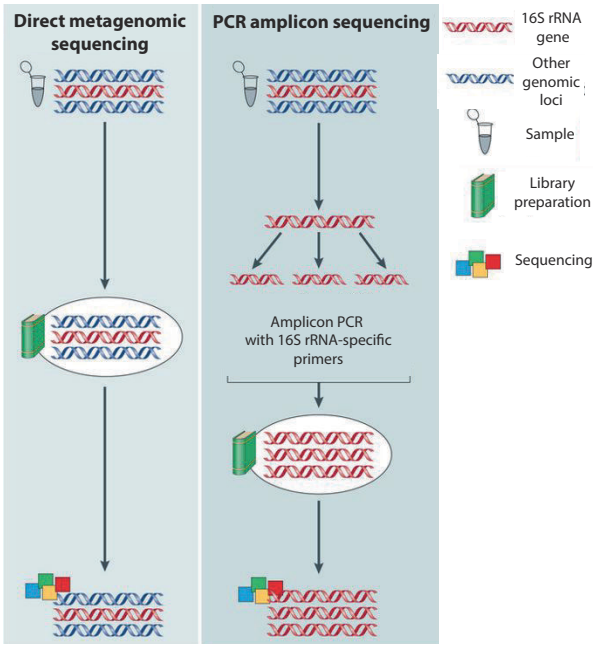


Figure 1.9: Sequencing approaches for microbial community analysis. Principal classes of sequencing techniques are targeted amplicon sequencing and untargeted Whole Genome Shotgun (or metagenomic) sequencing. While the former can cheaply sequence a single marker gene across a whole community, with the downside of potential amplification biases, the latter allows largely bias-free sequencing of the complete genomic content of all species (not restricted to a single marker gene). However, metagenomic sequencing provides less coverage per gene and is substantially more expensive than amplicon sequencing. Adapted with permission from Houldcroft et al. (2017).

16S sequences from all domains of life, were observed to not detect certain groups, for instance environmental isolates classified as *Actinobacteria* (Farris and Olson, 2007). This problem also affects large cooperative efforts, the Earth Microbiome Project (EMP) for instance switched its primer recommendation after initial results showed that the ubiquitous marine SAR11 clade, as well as environmentally significant *Crenarchaeota* and *Thaumarchaeota*, failed to be correctly amplified (Walters et al., 2016). Besides primer mismatches, biases of 16S amplicon sequencing also include polymerase errors, chimera formation and multi-template amplification biases (Tremblay et al., 2015), as well as the number of PCR cycles and amplicon size (Bonnet et al., 2002; Walker et al., 2015).

To circumvent amplification biases, investigators started applying Whole Genome Shotgun (WGS) sequencing methods directly to microbial samples (see Figure 1.9). While considerably more expensive than amplicon sequencing, WGS approaches measure not only a single genomic region (e.g. within the 16S rRNA gene) but instead yield sequence information on the full genetic content of a sample. This includes information of all community members, which inspired the term "metagenomics" (Handelsman, 2004). As a testament to its advantages, the first large metagenomic expedition, studying marine communities in the Sargasso Sea through WGS high-throughput sequencing, detected a total of 1800 bacterial and archaeal species (Venter et al., 2004). Findings included 148 previously unknown phylotypes, partially missed because of primer biases, and a wealth of novel functional genes.

Parallel innovations in computational assembly methods now also allow the routine reconstruction of functional genes, pathways and even whole genomes (metagenome-assembled genomes or MAGs) from metagenomic sequence data (Olson et al., 2017). Assembly of distinct genomes from the same sample became possible through the introduction of diverse binning techniques, for instance based on GC content. The usefulness of metagenomic assembly was recently highlighted through the reconstruction of over 800 bacterial draft genomes, assigned to 35 phyla newly proposed phyla, which constitute the candidate phyla radiation (CPR, Brown et al. (2015)).

Metagenomic approaches can furthermore increase the resolution of taxonomic classification through the use of multiple conserved marker genes (Sunagawa et al., 2013; Truong et al., 2015; see subsection 1.2.3) and additionally allow insights into the functional repertoire of a sample, which enables for instance functional comparisons between different communities (Mitra et al. (2011); see Figure 1.10). A highly similar protocol can be adapted for gene expression analysis, coining the term metatranscriptomics. In contrast to metagenomics, this variant measures the comprehensive activity (rather than abundance) of genes and thus provides a complementary view on microbial communities (Bashiardes et al. (2016); see Figure 1.10). However, unbiased metagenomic and metatranscriptomic sequencing comes at the price of overall reduced coverage of the 16S rRNA gene, preventing the detection of less abundant taxonomic groups (Gohl et al., 2016). These taxa can thus still only be economically recovered through amplicon sequencing (given appropriate primers).

Recently, a class of methods known under the umbrella of "third-generation" (or long-read) sequencing started reaching maturity (Bleidorn, 2016). In contrast to high-throughput approaches, which are restricted to read lengths of typically a few hundred base pairs, long-read sequencing methods produce reads with tens to hundreds of Kbp (kilo base pairs), albeit at much lower overall read numbers and with noticeably higher error rates (Goodwin et al., 2016). Examples of long-read technologies include Pacific Biosciences SMRT, Oxford Nanopore MinION and Illumina 10x Genomics (Goodwin et al., 2016). The drastically increased read length can strongly benefit genome assembly, in particular for rare species or in the presence of low-complexity regions (Sharon et al., 2015). It may furthermore substantially improve the recovery of whole operons and metabolic pathways (Tracanna et al., 2017)—even the sequencing of whole plasmids and small microbial genomes in one go is within reach.

Another major innovation in microbial sequencing was the advent of single-cell sequencing (Stepanauskas, 2012). While traditional primer-based amplification is restricted to specific genomic regions, novel polymerases isolated from phages allow the randomly distributed amplification of whole genomes (through multiple displacement amplification or MDA, Dean et al. (2002)). A typical workflow involves the use of cell

sorting technologies to isolate single cells of interest, whose genomes are subsequently amplified with MDA and sequenced through WGS sequencing. Since cells are sequenced individually, this methodology allows the cell-specific assembly of genomes, eliminating the need for complex and potentially error-prone computational binning. Single-cell technologies lead to the discovery and genome assembly of many novel and deeply branching taxonomic groups, including the DPANN (Rinke et al., 2013) and the *Asgardarchaeota* (Zaremba-Niedzwiedzka et al., 2017) superphyla.

A more recent sequencing approach is based on a combination of poly(A)-tailing and long-read sequencing, which allowed the recovery of a million full-length 16S rRNA with comparatively low error rates from a single study with only 19 samples (Karst et al., 2018). This approach can also reduce or avoid certain amplicon biases, including primer mismatches and chimeras, and may substantially expand reference databases of representative 16S rRNA sequences.

Prominent sequencing methods face the major problem that the achieved sequencing depth (within certain limits) is inherently arbitrary, depending solely on technical factors, such as the amount of DNA extracted or sequenced, as well as fluctuations in read quality (Faust and Raes, 2012). The counts generated by sequencing are thereby compositions, which cannot be treated as absolute abundances, thereby precluding the direct comparison of samples and complicating many analyses (see subsection 1.3.2). Approaches utilizing flow cytometry, DNA spike-in, FISH or quantitative PCR (qPCR) have been proposed for calibrating read counts and reporting values closer to absolute abundances, with promising initial results (Gifford et al., 2011; Nakatsuji et al., 2013; Props et al., 2017; Vandeputte et al., 2017a).

As previously mentioned, an important step in microbial sequence analysis is the DNA extraction step. With the analysis of more exotic habitats and communities, it is increasingly recognized that standard extraction protocols may systematically miss important groups that fail to be lysed, caused for instance by differences in cell wall chemistry (Kennedy et al., 2014; Salonen et al., 2010). Similarly, sample storage conditions may impact downstream analyses (Cardona et al., 2012; Vandeputte et al., 2017b) and contaminants from extraction kits (the so-called "kitome") or PCR reagents may strongly confound analyses of low-biomass samples (Salter et al., 2014).

1.2.3 Bioinformatic preprocessing and phylotypes

After sample collection and sequencing, categorization of reads into phylotypes, in order to obtain relative abundance profiles, is the first step of every computational microbiome analysis (Caporaso et al., 2010; Schloss et al., 2009). This necessitates the preprocessing and cleaning of raw sequencing data, which includes demultiplexing, removal of adaptor and barcode sequences, trimming or removal of low-quality reads,

joining of paired-end reads, denoising and chimera removal.

After this data cleaning process, abundances for taxa and/or phylotypes can be computed, followed by further downstream analyses such as diversity computation and visualization. All-in-one bioinformatic software pipelines, such as mothur (Schloss et al., 2009) and QIIME (Caporaso et al., 2010), have been developed to facilitate this complex workflow. These highly successful frameworks aim to provide reproducible and flexible solutions for typical microbiome analysis workflows, unlocking many options for sample analysis to microbiologists with more limited bioinformatic expertise. If metagenomic data is available, specialized pipelines such as ngless (Coelho et al., 2018), HUMAnN2 (Franzosa et al., 2018) and bioBakery (McIver et al., 2017) can be used to additionally infer abundances for functional gene categories and pathways, for instance by mapping metagenomic reads against the KEGG database (Kanehisa et al., 2017).

To obtain the abundance profile of a sample, the simplest method is taxonomic classification, in which sample reads are mapped to large 16S rRNA reference databases, such as the RDP database (Cole et al., 2014), Greengenes (DeSantis et al., 2006), SILVA (Yilmaz et al., 2013) or MAPref (Matias Rodrigues et al., 2017), which include up to millions of full-length 16S rRNA sequences of cultured and uncultured organisms. These reference sequences are annotated with taxonomic lineages from various sources (Balvočiūtė and Huson, 2017), commonly the NCBI taxonomy database (Federhen, 2011), and linked by the mapping process to sample reads, which results in taxon abundance profiles.

For the mapping step, either general-purpose tools, such as BLASTN (Camacho et al., 2009) and USEARCH (Edgar, 2010), or software optimized for 16S rRNA, such as the RDP Classifier (Wang et al., 2007) and MAPseq (Matias Rodrigues et al., 2017), are commonly used. The latter two utilize additional rank-specific confidences to increase accuracy and allow for partial lineage mapping. For metagenomic samples, which include sequences across all genes, approaches that map metagenomic reads to reference genomes can result in increased taxonomic resolution and avoid the previously mentioned shortcomings of marker genes and amplicon sequencing (see subsections 1.1.3 and 1.2.2). Methods for metagenomic classification include for instance the mOTU tool (Sunagawa et al., 2013) and MetaPhlAn2 (Truong et al., 2015).

While taxonomic mapping has the advantage of generating abundances for commonly used names, facilitating comparisons between studies, this approach only classifies reads that are sufficiently similar to previously observed 16S rRNA reference sequences and thus leads to the omission of sequences from new taxa (Schloss and Westcott, 2011). Since the vast majority of microbial diversity known from sequence-based approaches is thus far uncharacterized, taxonomic mapping of any form thereby

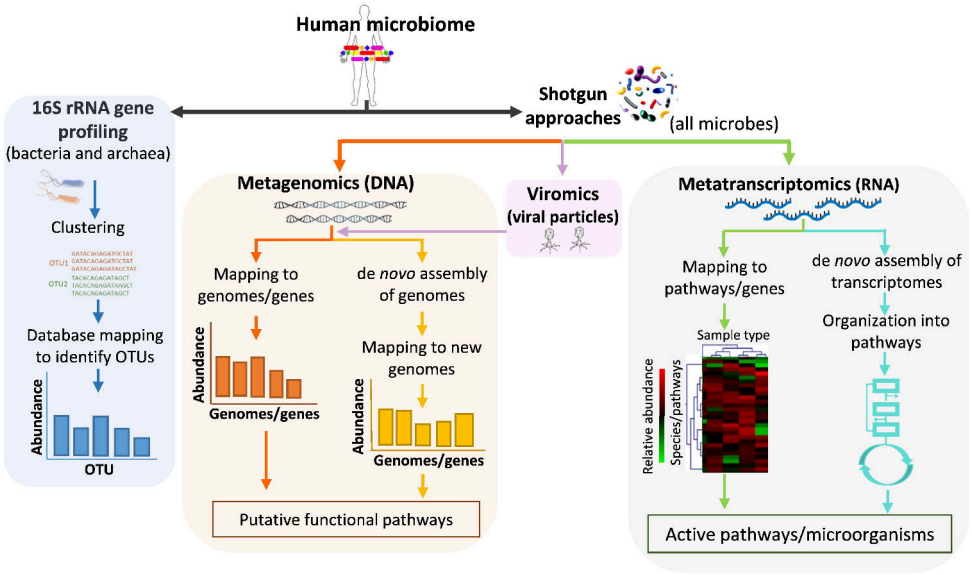


Figure 1.10: Bioinformatic workflows for microbial community analysis. Marker gene-based profiling (typically utilizing the 16S rRNA gene) allows the generation and comparison of abundance profiles for either taxa or phylotypes (OTUs, ESVs), which results in a coarse-grained overview of the community (here a human microbiome). Metagenomic and -transcriptomic approaches help to improve the resolution of taxa and phylotypes by providing information on more genes, while additionally enabling assembly of pathways and whole genomes. Metatranscriptomics furthermore provides insights into not only the presence but also the activity of particular pathways within a community. Credit due to Bikel et al. (2015).

typically results in large fractions of unmapped reads[15], even in samples from highly-studied environments, such as the human gut. Important insights on undescribed taxa, which may impact study conclusions and are often the most interesting targets in a study (particularly in environmental microbiology), can thereby be missed.

One way to address this issue is the use of operational taxonomic units (OTUs), which are clusters of 16S rRNA sequences with sufficiently high sequence identity to be operationally called the same unit of biodiversity. As previously described (see subsection 1.1.3), DNA-DNA hybridization methods lead to the suggestion of 16S rRNA identity thresholds that roughly correspond to species level delineations, for instance the still commonly used 97% cutoff or the more recently proposed 98.7% threshold (Stackebrandt, 2006). These cutoffs additionally ensure that the effect of sequencing errors are mitigated, which may otherwise strongly inflate the observed biodiversity (Kunin et al., 2010).

Commonly used software tools for OTU clustering include heuristic methods, such as UCLUST (Edgar, 2010), CD-HIT (Fu et al., 2012), UPARSE (Edgar, 2013) and Swarm (Mahé et al., 2014), or exact hierarchical clustering methods, such as HPC-CLUST (Matias Rodrigues and Mering, 2014). While generally more computationally expensive, exact methods tend to produce more robust and ecologically consistent clusters (Schmidt et al., 2014; Schmidt et al., 2015), while furthermore generating an ultrametric hierarchy of OTUs that can be treated as an unsupervised hierarchical

[15] unless specifically dealt with (Sunagawa et al., 2013)

taxonomy of OTUs (Matias Rodrigues et al., 2017), complementing classic taxonomies.

To infer OTUs abundances from samples (see Figure 1.10), several methods can be used. The *de novo* clustering approach involves clustering of all reads in one or more samples from scratch, thus utilizing full read information, which results in a comprehensive set of OTUs (Schmidt et al., 2015; Westcott and Schloss, 2015). However, this method is computationally demanding, as it scales quadratically in the number of input sequences, restricting it typically to small sets of samples or necessitating the use of cheaper heuristics, which can introduce inconsistencies (Schmidt et al., 2015). Alternatively, *de novo* clustering can be performed on a per-sample basis, which allows each sample to be efficiently analysed in a parallel fashion, but precludes direct comparisons between samples. For the same reason, different data sets whose OTUs were clustered independently also cannot be compared.

An alternative approach that aims at mitigating these problems is the closed-reference approach: sequences in a reference 16S rRNA database are clustered into OTUs, and reads from samples can subsequently be mapped against this database (Westcott and Schloss, 2015). This method mirrors taxonomic classification, but results in OTU abundances rather than taxon abundances. The computational advantage of the closed-reference approach is that clustering needs to be done only once (or infrequently, during database updates), after which mapping can be done efficiently for large numbers of samples and reads. Another crucial advantage compared to *de novo* clustering is that inter-sample comparability is ensured, since all samples are mapped against the same reference database. This also extends to samples from different studies, even if these used different hypervariable regions for sequencing, since reference databases consist of full-length 16S rRNA sequences covering all regions. However, samples mapped to different databases cannot be compared, since they are based on different references.

Despite all advantages, the closed-reference approach shares the defining disadvantage of taxonomic mapping: new sequences with insufficient similarity to reference OTUs cannot be mapped and these reads are thereby lost for analysis. Unmapped reads can make up a substantial fraction of any given sample, in particular when analyzing more exotic habitats, and may thus lead to an incomplete picture of the underlying community. Furthermore, accurate identity thresholds cannot be guaranteed when mapping reads to a database, which can be illustrated as follows: it is possible that two reads match the same database sequence with identity higher than the OTU cutoff (for instance 97%), classifying them as the same OTU, while the identity between these reads may be much lower (down to 94% in this case), which would more correctly put them into different OTUs (Westcott and Schloss, 2015). As another caveat, confident assignments may be impossible if multiple reference sequences, assigned to different OTUs, are identical within the hypervariable region used to sequence the input sample

(Westcott and Schloss, 2015).

To reduce the problem of unassigned reads, the open-reference approach has been proposed, a hybrid algorithm that combines the advantages of closed-reference mapping and *de novo* clustering (Westcott and Schloss, 2015). This method starts by performing the full closed-reference workflow, resulting in abundances for all OTUs found in the used reference database, as well as a set of reads that could not be confidently assigned. Next, these unmapped reads are clustered *de novo* to obtain abundances for OTUs not present in the reference database, leading to a more complete characterization of the sample community. The open-reference approach thus represents a tractable middle-ground between closed-reference and full *de novo* clustering, producing more comprehensive profiles than the former, while reducing the computational cost of the latter.

However, the method retains central disadvantages of *de novo* clustering: the second step may still be prohibitively expensive when facing large numbers of unassigned reads, data sets with independently clustered OTUs remain incomparable, and samples sequenced using different hypervariable regions cannot be consistently clustered, because reads from the same organism may cluster into different OTUs (depending on the region). Another disadvantage of the open-reference approach is that it combines two distinct OTU definitions, which effectively correspond to different thresholds and can thereby make interpretation more difficult (Westcott and Schloss, 2015).

An issue affecting all OTU assignment methods is that popular 16S rRNA identity cutoffs, such as the previously mentioned 97% or 98.7% thresholds for species distinction, have been calibrated using full-length sequences, while amplicon sequencing targets only a restricted region of the 16S rRNA gene. Different regions of this gene have been shown to evolve at different speeds and appropriate region-specific thresholds should therefore be used, which however are yet to be determined (Schloss, 2010).

Independent of which method is used for OTU assignment, considerable discrepancies between tools have been reported, regarding stability, cluster quality, consistency and sensitivity to parameter choice (Schloss and Westcott, 2011; Schmidt et al., 2014; Schmidt et al., 2015; Westcott and Schloss, 2015). While average-linkage clustering typically ranked amongst the best-performing OTU clustering methods in these benchmarks, and thus appears to be a solid default choice, care is therefore advised.

OTU-based approaches have a successful history in microbial ecology (Arumugam et al., 2011; Chaffron et al., 2010; Fierer et al., 2010; Huttenhower et al., 2012; Sunagawa et al., 2015) and remain widely used, but more recently, methods aiming at detecting finer-scale ecological structure have been developed. Examples of these tools include MED (Eren et al., 2015), DADA2 (Callahan et al., 2016), Deblur (Amir et al., 2017) and UNOISE (Edgar and Flyvbjerg, 2015), which use information theoretical approaches

or statistical sequencing error models to remove likely erroneous sequence variants. Otherwise, these tools retain full sequence information and thus allow the distinction of phylotypes down to single nucleotide differences (Callahan et al., 2017). Depending on the method, these phylotypes have been termed exact/amplicon sequencing variants (ESVs/ASVs), tag sequences, sOTUs, oligotypes or 100% OTUs, but will for simplicity be referred to as ESVs in this thesis.

The increase in resolution provided by ESVs has in some cases uncovered distinct ecological patterns not detectable (or less pronounced) at the OTU level, such as season-specific variations of SAR11 ESVs and increased temporal correlation between these ESVs and specific bacteriophages (Needham et al., 2017). ESVs have furthermore been used in a high-level analysis of the Earth Microbiome Project data set, which allowed the detection of marked habitat specificity at the ESV level, a signal that was considerably weaker in 97% OTUs and higher taxa (Thompson et al., 2017). Further advantages include, that these approaches use the full read information of a sample, are independent of reference databases and can computationally scale to large numbers of samples (Callahan et al., 2017).

However, while samples using the same hypervariable region can be compared via ESVs, combining different regions in the same data set still requires a closed-reference approach. This problem may be alleviated in the future by the previously mentioned sequencing method proposed by Karst et al (Karst et al. (2018); see subsection 1.2.2), which could effectively allow the assignment of full-length ESVs to circumvent the issue of hypervariable regions (apart from other advantages, such as primer bias mitigation). Another disadvantage of ESV-based approaches is that they may provide a too fine scale for certain scientific questions. For instance, ecologically meaningful patterns may be strongest at higher granularity, and could thus be significantly weakened or even lost when pattern-specific OTUs are split into smaller ESVs, each carrying only part of the signal. The strong increase in data set dimensionality induced by ESVs may furthermore pose a considerable challenge to studies that are hard-pressed for statistical power, such as metagenome-wide association studies (MGWAS, Wang and Jia (2016)), due to substantially increased numbers of significance tests and thus stricter multiple testing correction. Additional shortcomings include, that higher dimensionality can increase the likelihood of overfitting statistical models[16] and that it may prevent the use of methods that require full-rank input matrices (with higher numbers of samples than variables).

Both traditional OTUs and ESVs are also limited by fundamental evolutionary constraints of the sequenced marker gene (e.g. 16S rRNA), which may lead to insufficient resolution for distinguishing ecologically meaningful units. The effect of this was recently observed in *Microcystis*, for which certain ESVs were shown to

[16] the so-called "curse of dimensionality" (Bellman, 1961)

correlate with toxicity in a freshwater environment, but follow-up analysis with culture strains assigned to the same ESVs revealed poor correspondence in terms of toxicity phenotype (Berry et al., 2017).

If multiple genes were sequenced, or metagenomic data is available, other approaches for the detection of fine-scale patterns can thus be applied to increase resolution beyond a single marker gene. As previously discussed, MLST approaches (based typically on 6-8 marker genes) have been used to distinguish even highly related strains, which may—despite their phylogenetic similarity—display important phenotype differences, for instance in the case of pathogenic vs non-pathogenic strains (Madigan et al., 2014). Recently, a number of more resolved strain detection approaches have been developed, such as metaSNV (Costea et al., 2017), StrainPhlAn (Truong et al., 2017) and DESMAN (Quince et al., 2017), which use information from metagenomic samples to identify fine-scale biodiversity units at the strain level via single nucleotide variations (SNVs) (Segata, 2018). These tools either map input sequences to reference genomes, an approach taken by metaSNV and StrainPhlAn, or work with *de novo*-assembled contigs in a reference-free approach (DESMAN). The increased resolution compared to single-gene methods was for instance shown to improve distinction of pathogenic and commensal *E. coli* strains (Ward et al., 2016), uncovered subject-specific temporally stable strain patterns in human gut and oral microbiomes (Donati et al., 2016; Truong et al., 2017), and allowed the tracking of specific engrafted strains after fecal transplantation (Li et al., 2016).

Field-wide challenges are becoming popular in an increasing number of scientific disciplines, as for instance in protein structure prediction (CASP, Moult et al. (2016)) and gene regulatory network inference (DREAM, Stolovitzky et al. (2007)), with the aim of consistently benchmarking state-of-the-art computational prediction methods. Following this trend, The Critical Assessment of Metagenome Interpretation (CAMI, Sczyrba et al. (2017)) challenge is a recent concerted effort to foster robustness and reliability in metagenomics software. The CAMI challenge recently completed its first round with 19 participating teams, competing in benchmarks for taxonomic classification, binning and genome assembly. Results showed generally robust performance for taxonomic classification of higher ranks (family and above) across tools, but considerable drops in accuracy at the genus, species and strain levels. Furthermore, this evaluation highlighted widespread difficulties in assembling and binning closely related genomes (Sczyrba et al., 2017).

1.2.4 Databases for microbial ecology

Billions of sequences from metagenomic and amplicon sequencing samples have been deposited at dedicated nucleotide databases, such as the EBI European Nucleotide

Archive (ENA, Harrison et al. (2018)) and the NCBI Sequence Read Archive (SRA, Leinonen et al. (2010)). These resources now provide raw reads for more than a million sequencing analyses of microbial samples (1'214'604 in the SRA, accessed December 2018), which cover highly diverse environments across all continents and oceans, ranging from clinically or agriculturally relevant animal-associated microbiota to a variety of extreme and exotic habitats (see subsection 1.1.1). In conjunction with rapidly growing reference databases, raw sequence collections provide an exceptional foundation for recent efforts to provide detailed phylotype abundance and metadata information within comprehensive, integrated resources. These microbial ecology-oriented databases aim at making the wealth of raw sequence data amenable to end users for ecological analysis, for instance in the context of cross-biome analyses or for the investigation of specific questions and hypotheses.

Current resources include EBI Metagenomics (now MGnify, Mitchell et al. (2018)), QIITA (Gonzalez et al., 2018), MG-RAST (Keegan et al., 2016), IMNGS (Lagkouvardos et al., 2016) and curatedMetagenomicData (Pasolli et al., 2017), which cover a wide variety of data types and analysis workflows. While all of these databases allow users to track biodiversity units (either phylotypes or recognized taxa) across studies and environments, considerable differences in scope exist between them. For instance, there is substantial variation in terms of numbers of publicly available (and consistently processed) samples, the extent and type of on-page analyses and webpage interactivity, and to what level functional analysis of metagenomic data is possible. MGnify, QIITA and IMNGS all aim at providing large numbers of comparable samples, processed for the 16S-based tracking of phylotypes or taxa across studies, and encourage (MGnify), enforce (QIITA) or don't explicitly consider (IMNGS) standardized MIxS sample checklists (Yilmaz et al., 2011) for metadata annotations. In contrast, curatedMetagenomicData features a considerably smaller quantity of samples but employs higher levels of human curation and metadata standardization. MG-RAST also features smaller numbers of publicly available data sets and is less focused on taxa or phylotype tracing, but instead provides extensive capabilities for functional mapping and analysis, with a more reference genome-centric perspective[17].

The Microbe Atlas Project database (MAPdb, Rodrigues et al. (manuscript in preparation)) is a new resource, currently in development, that focuses on extensive interactivity and exploration. It provides a substantially more comprehensive collection of comparable microbial sequencing samples than the previously mentioned databases: 1'018'489 consistently processed samples are currently available in the pre-release version of MAPdb, compared to 83'646 in MGnify (processed with the latest pipeline, version 4.1) and 79'568 in QIITA (processed in closed-reference mode) (all databases accessed December 2018). Its main strengths lie in environmental exploratory

[17] MGnify also features increasing capabilities for functional analysis

discovery and hypothesis generation, for instance through facilitated investigation of environmental niches (via unsupervised sample clusters) and an array of options for users to compare their own samples to a global background.

Phylotype abundance databases are complemented by genome resources, such as IMG/M (Chen et al., 2017), which provide access to assembled genomes from a variety of sources, for instance culture strains, environmental MAGs or single cell genomes (which vary in completeness and quality). Through these genome collections, it becomes possible to link phylotypes of interest (identified in the previously mentioned databases) with additional information, albeit this requires that the corresponding organism was previously sequenced and that the genome includes the phylotype-specific marker gene. Another important resource for valuable complementary information are phenotypic traits, such as oxygen requirements and antibiotic resistance, which may be inferred by genomic or taxonomic mapping, or alternatively via increasingly available genome-scale metabolic mathematical models (Cuevas et al., 2016). Mappings of phenotypic traits and metabolic models are provided by database resources such as PATRIC (Wattam et al., 2017) and ModelSEED (DeJongh et al., 2007) and can be linked to phylotypes, in order to inform more detailed follow-up analysis of ecologically interesting patterns detected in exploratory databases.

While several community-driven resources exist for ecological interactions (for instance DRYAD[18] and IWDB[19]), these mostly include interactions between multicellular organisms, while dedicated databases for curated microbial interactions are currently lacking.

1.2.5 Standard methods for microbiome analysis

With the advent of sequence-based sampling of microbial communities—and the consequent increase in sampling comprehensiveness—diversity-based methods from classic ecology (Whittaker, 1972) became heavily used tools in microbiome research (Goodrich et al., 2014; Lozupone and Knight, 2008). These methods allowed for instance the better quantification of community-level differences between health- and disease-associated microbiota (Pascal et al., 2017; Zhang et al., 2018b), different hosts (Ley et al., 2008) and environments (Thompson et al., 2017) (including environmental gradients (Sunagawa et al., 2015)).

Ecological theory broadly distinguishes diversity into alpha, beta and gamma diversity (Whittaker, 1972), with the former two being the prime focus in microbial ecology research (Lozupone and Knight, 2008). Alpha diversity describes the local community diversity per patch or sample, which in microbial ecology typically means

[18] `http://datadryad.org/`
[19] `https://www.nceas.ucsb.edu/interactionweb/`

a sequencing sample (see Figure 1.11 A). The simplest alpha diversity index is species richness, i.e. the number of observed species in a sample, which assumes that species delineations are meaningful and furthermore disregards abundances. In order to also incorporate relative species abundances and measure for instance whether a community is dominated by a few taxa, species evenness can be used (commonly measured through the Shannon index (Vajda et al., 1950)). Species evenness can be informative if disturbances have a stronger impact on abundance than on presence/absence, for example in contaminated environments (Forster et al., 2018).

Since microbial species delineations are largely arbitrary (see subsection 1.1.3), phylogenetic indices for species richness (Faith's Phylogenetic Diversity, Faith (1992)) and evenness (θ, Martin (2002)) have been developed, which use the leaf-leaf distance between species in a phylogenetic tree to weight diversity according to evolutionary divergence. For instance, if two samples have the same species richness, but one includes very divergent species, while the other features species close to the operational microbial species cutoff, the former would be assigned a higher diversity by these indices.

A problem that any ecological diversity study faces are sampling biases. Usually, only a limited number of patches can be sampled within any given environment, due to lack of resources, which can hinder detection of less abundant species. Importantly, this problem also precludes the accurate comparison of alpha diversities between different environments with different sampling coverages, even if the absolute number of observed specimen is the same (or has been post-processed to account for sampling effort) (Chao and Jost, 2012).

A large theoretical body on estimating total sample diversity has been developed to tackle this problem, yielding diverse parametric and nonparametric methods. The nonparametric Chao estimator (Chao, 1984), for instance, uses the fraction of singleton and doubleton species to estimate the number of unobserved species and is currently the most widely used index for approximating total species richness. This concept has recently been generalized via Hill numbers to phylogenetic and functional diversity definitions, from which also indices for evenness and other abundance weightings can be derived (Chao et al., 2014). The framework can furthermore be used for diversity interpolation and extrapolation, which allows the equalization of coverage between samples and thus enables valid comparisons of diversity (Chao et al., 2014).

Alpha diversity analysis has for instance revealed reduced diversity in various disease conditions associated with dysbiosis (DeGruttola et al., 2016; Mosca et al., 2016) or in connection with host lifestyle factors, such as smoking (Feigelman et al., 2017). Reduced alpha diversity was furthermore found in contaminated environments (Forster et al., 2018), possibly indicative of reduced ecosystem status and functioning.

While alpha diversity is concerned with the diversity of a single sample, beta

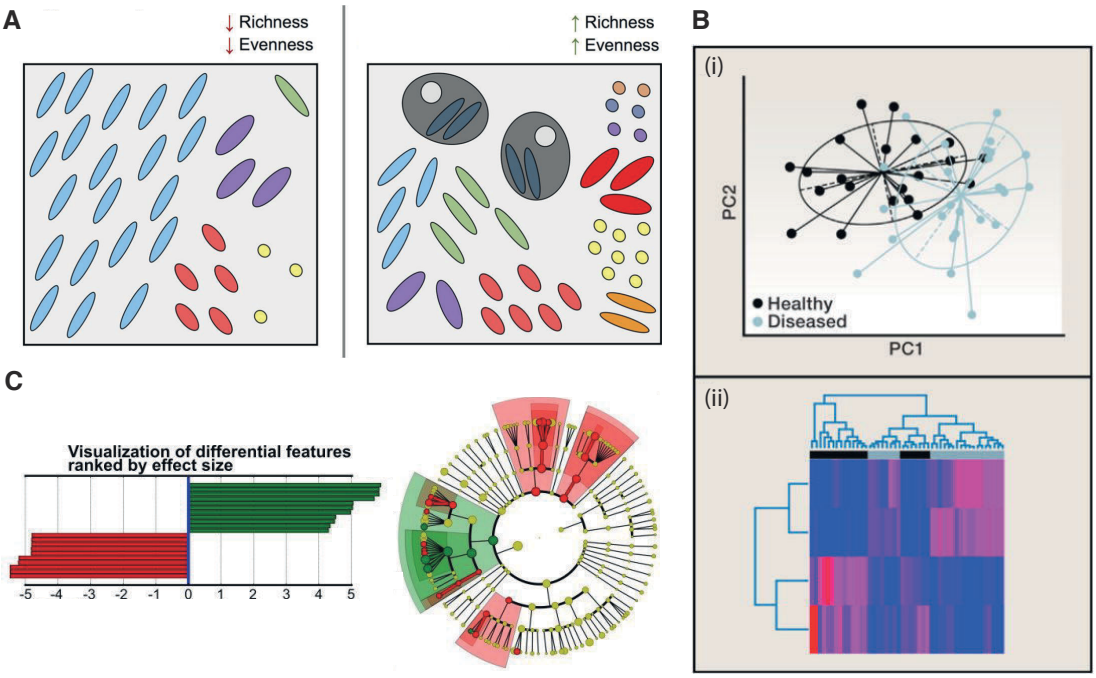


Figure 1.11: Standard types of analyses for microbial abundance data. **A** General classes of alpha-diversity indices: while richness quantifies the overall number of species in a community, evenness also takes abundances into account and results in lower diversity estimates for communities that are numerically dominated by one or few species. Adapted with permission from Laforest-Lapointe and Arrieta (2018). **B** Beta diversity indices allow the quantification of similarity between different communities, which can subsequently be visualized via ordination plots or cluster dendrograms. Statistical tests should furthermore be applied to test the significance of systematic community differences between groups (e.g. healthy vs. diseased). Adapted with permission from Goodrich et al. (2014). **C** Differential abundance analysis allows the detection of species that are significantly differentially abundant between conditions (i.e. microbial biomarkers). These may for instance play a role in disease progression or show specific habitat preferences. Adapted with permission from Segata et al. (2011).

diversity quantifies community variation across different samples (Whittaker (1972); see Figure 1.11 B). A classic species-based beta diversity index is the unweighted Jaccard index, which measures the fraction of overlapping species over the union of all species between two samples (Jaccard, 1901). The weighted Jaccard (Chao et al., 2004) and Bray-Curtis dissimilarity (Bray and Curtis, 1957) indices additionally account for abundance differences between overlapping species.

Similar to species-based alpha diversity indices, these beta diversity indices however face the problem of arbitrary species delineations. This problem is typically addressed through the use of phylogenetically informed indices, such as the weighted and unweighted Unifrac indices (Lozupone and Knight, 2005; Lozupone et al., 2007). Furthermore, the weighted and unweighted Taxa INteraction-Adjusted indices (TINA) were recently proposed, which consider the ecological structure of communities in order to achieve a more fine-scale detection of community differences (Schmidt et al., 2017).

After computation of pairwise sample distances or dissimilarities, results are typically visualized via clustering or ordination approaches, the latter including for instance Principal Coordinates Analysis (PCoA) or non-metric multidimensional scaling (NMDS) (Goodrich et al., 2014; Paliy and Shankar, 2016; see Figure 1.11 B). Differences

between groups (e.g. conditions) within the distance matrix can furthermore be tested for statistical significance, using for example the permutational analysis of variance (PERMANOVA, Anderson (2001)) test.

Beta diversity analyses have been used to answer a variety of questions, for instance to identify community-level differences between human body sites (Huttenhower et al., 2012) and diets of animal hosts (Ley et al., 2008), or to match the skin microbiota of human subjects to touched objects (Fierer et al., 2010).

Apart from community-level differences between conditions, investigators are in many cases also interested in single taxa that are differentially abundant (microbial biomarkers), for instance in the context of MGWAS studies (Wang and Jia, 2016) (see Figure 1.11 C). While simple statistical hypothesis tests, such as the two sample t-test or the Mann-Whitney U test, have been used to this end, several potential shortcomings of these straightforward approaches have led to the development of new methods for differential abundance detection.

For instance, methods from the differential gene expression field have been proposed to also be effective for microbial abundances (McMurdie and Holmes, 2014). Tools such as edgeR (Robinson et al., 2010), DESeq2 (Love et al., 2014) and Voom (Law et al., 2014) employ parametric approaches that use specific distributional assumptions to increase sensitivity compared to non-parametric tests, which makes them more suited for smaller data sets (given that assumptions are met). These methods furthermore employ optimized normalization procedures, which aim at ensuring comparability between samples.

Apart from differential expression methods, microbiome-specific methods have also been developed, such as LEfSe (Segata et al., 2011) and metagenomeSeq (Paulson et al., 2013). While LEfSe uses subgroup consistency information to increase sensitivity, metagenomeSeq employs a custom normalization technique, combined with an undersampling-aware statistical model, to address microbiome-specific issues. Other approaches motivated by considerations from the field of compositional mathematics, such as ANCOM (Mandal et al., 2015) and ALDEx2 (Fernandes et al., 2014), provide theoretically sound ways of tackling the compositionality problem, which is expected to be severe in microbial sequencing samples (Gloor et al. (2017); see subsection 1.3.2). ANCOM-II additionally accounts for the effects of excess zeros in compositional data sets (Kaul et al., 2017).

Evaluation studies comparing a variety of differential abundance methods in a range of simulated data sets generally found noticeable differences between tools, depending on data characteristics and simulation setup. While parametric approaches developed for differential gene expression showed promising performance in some studies (McMurdie and Holmes, 2014; Weiss et al., 2017), typical assumptions of these methods may be regularly violated in microbiome data, with varying effects on

sensitivity and false discovery rate (FDR) (Hawinkel et al., 2017; Weiss et al., 2017). While a recent benchmark study found that correlations between OTUs can drastically increase false positive predictions across methods (Hawinkel et al., 2017), differential abundance approaches that account for inter-species correlations are still rare (notable exceptions are GLL (Aliferis et al., 2010a) and DAA (Menon et al., 2018); see also Figure 1.15 C). Albeit no interaction-aware methods were tested by Hawinkel et al, GLL was successfully applied in other microbiome settings (Statnikov et al., 2013a; Tackmann et al., 2018). Some of the mentioned tools, such as LEfSe, can also detect differentially abundant functional genes or pathways (Segata et al., 2011).

After initial promising benchmarks on a variety of tasks (Knights et al., 2011a; Statnikov et al., 2013b), machine learning methods are now being increasingly used for microbiome analyses. The distinguishing property of these algorithms is their ability to detect complex patterns in multi-dimensional data and their potential to continually increase performance as data grows (Kodratoff and Michalski, 2014). Applications of supervised classification methods cover diverse applied areas in need of high accuracy, such as disease prediction (Pasolli et al., 2016; Zeller et al., 2014), source tracking (Knights et al., 2011b), detection of organic contamination (Cordier et al., 2018) and correction of mislabeling errors (Knights et al., 2011c). But these methods have also been used to investigate study-specific questions more deeply, such as temperature impact (Sunagawa et al., 2015), the importance of OTUs vs. environmental factors in predicting OTU abundances (Lima-Mendez et al., 2015), and the validation of enterotypes (Arumugam et al., 2011).

Apart from making accurate predictions, supervised classification methods can furthermore be used for feature selection, which allows the identification of highly predictive OTUs, environmental variables and functional genetic markers (Knights et al., 2011a; Pasolli et al., 2016; Zeller et al., 2014). Feature selection may for instance improve classification performance, but most importantly can provide further insights into important relationships between discovered markers and the phenomenon of study ("knowledge discovery", Heckerman (1997)).

A widely-used machine learning method in microbiome studies, with excellent robustness to overfitting and intrinsic feature selection capabilities, are Random Forests (Breiman, 2001). This supervised classification and regression framework trains large numbers of decision trees on random subsets of features and observations, and subsequent predictions are based on combined votes across trees. Classifiers of this class were constantly among the best-performing methods in a variety of microbiome-related classification tasks, including disease prediction, human body site identification and the forensic matching of the skin microbiome of subjects to computer mice or keyboards they used (Knights et al., 2011a; Pasolli et al., 2016; Statnikov et al., 2013b).

1.3 Computational inference of microbial ecosystem structure

"Wie alles sich zum Ganzen webt,
Eins in dem andern wirkt und lebt!"[20]
—Dr. Heinrich Faust (*Faust, Part One*[21])

1.3.1 Current tools for microbial interaction prediction

The previously discussed methods (see subsection 1.2.5) are mainly descriptive in nature and thus limited to comparatively modest (albeit important) questions, for instance regarding the species content and diversity of a community or whether whole communities and single taxa differ across conditions. However, microbiota are more than mere sets of species to count and compare: they represent complex ecosystems of interacting biotic and abiotic components (see subsection 1.1.1) and these interactions ultimately create the patterns we observe in sequencing snapshots or under the microscope. In order to truly understand how microbial communities are shaped by their members and extrinsic factors, and how they impact their environment (e.g. their host), a structural and mechanistic understanding of interactions is thus pivotal. Only then can detailed models be proposed, validated and studied to gain deeper insights into emergent system properties (see subsection 1.1.2), to accurately predict the impact of perturbations and to generate richer hypotheses that fuel the scientific process.

As mentioned in subsection 1.1.2, the sheer amount of microbial species, as well as the plenitude of modulating environmental conditions and the technical difficulties of accurately measuring interactions, make the experimental validation of microbial interactions only possible for trivial numbers of species pairs and conditions. While high-throughput co-culturing approaches may reduce this problem in the future (see subsection 1.2.1), computational methods thus remain the only means for predicting interactions within realistically-sized microbial ecosystems.

Several different classes of computational methods have been developed, each of which exploits different types of data and signals to infer ecological relationships, and thus exhibits different strengths and limitations. The first type of tools uses temporal information from longitudinal microbial abundance measurements to predict interactions, based on time-lagged shifts in the abundance of a species in response

[20] *"How all things interweave as one*
and work and live each in the other!"
[21] Written by *Johann Wolfgang von Goethe*

to earlier abundance shifts in others (see Figure 1.12 A). Generalized Lotka-Volterra (gLV) models are a popular framework in this context, for instance used by MDSINE (Bucci et al., 2016), MetaMIS (Shaw et al., 2016), LIMITS (Fisher and Mehta, 2014) and the approach used in (Stein et al., 2013). These ordinary differential equation models are typically fitted to temporal data via autoregressive techniques and predict signed, weighted and asymmetric interactions. In addition to structural predictions, gLV models furthermore allow the simulation of temporal ecosystem dynamics. Other methods employ the autoregressive integrated moving average (ARIMA) process (Ridenhour et al., 2017) or Granger causality (Gibbons et al., 2017) to infer directed interactions between microbes from longitudinal data. eLSA (Xia et al., 2011) uses local similarity search to match temporal windows and thus detect time-delayed linear interactions between microbes, while also accounting for information provided by technical replicates. Dynamic Bayesian Networks (DBNs), a probabilistic approach that excels at handling uncertainty and noisy data, have also been used successfully to infer structure and dynamics from microbial time-series data, for instance within the infant gut microbiome (McGeachie et al., 2016). Temporal methods have furthermore detected substantial competition within mouse gut communities (Stein et al., 2013), revealed multiple dynamic regimes in the human gut[22] (Gibbons et al., 2017), and showed the importance of *Flavobacteria* for seasonal succession dynamics in coastal microbial ecosystems (Pollet et al., 2018).

A noteworthy challenge of temporal methods is the correct setting of sampling intervals, as some patterns may only be detectable at certain time resolutions (which can furthermore be tool-dependent) (Cao et al., 2017). Additionally, perturbations may be needed to make temporal data sufficiently informative and, furthermore, the identification of correct structures may generally be impossible in certain scenarios (Angulo et al., 2017; Cao et al., 2017). The distinction of interaction-conveyed effects from stochastic effects also requires particular care (Faust et al., 2018). In addition, though longitudinal data is becoming increasingly available for some heavily studied ecosystems, such as mammal gut or aquatic environments (Caporaso et al., 2011; Dam et al., 2016; David et al., 2014b; Gilbert et al., 2012), data quantities are still considerably lacking behind cross-sectional data. This includes both number and heterogeneity (e.g. subjects, conditions, environments) of studies, which prevents the longitudinal analysis of many interesting, more exotic environments. Comprehensive comparative benchmarks of temporal methods are currently lacking, and in consequence, method-specific biases remain largely unknown, which impedes informed tool choice.

Inspired by the availability of increasing numbers of genomes and MAGs, methods utilizing metabolic complementarity and redundancy between different species have

[22] one driven mostly by environmental and the other by intrinsic effects

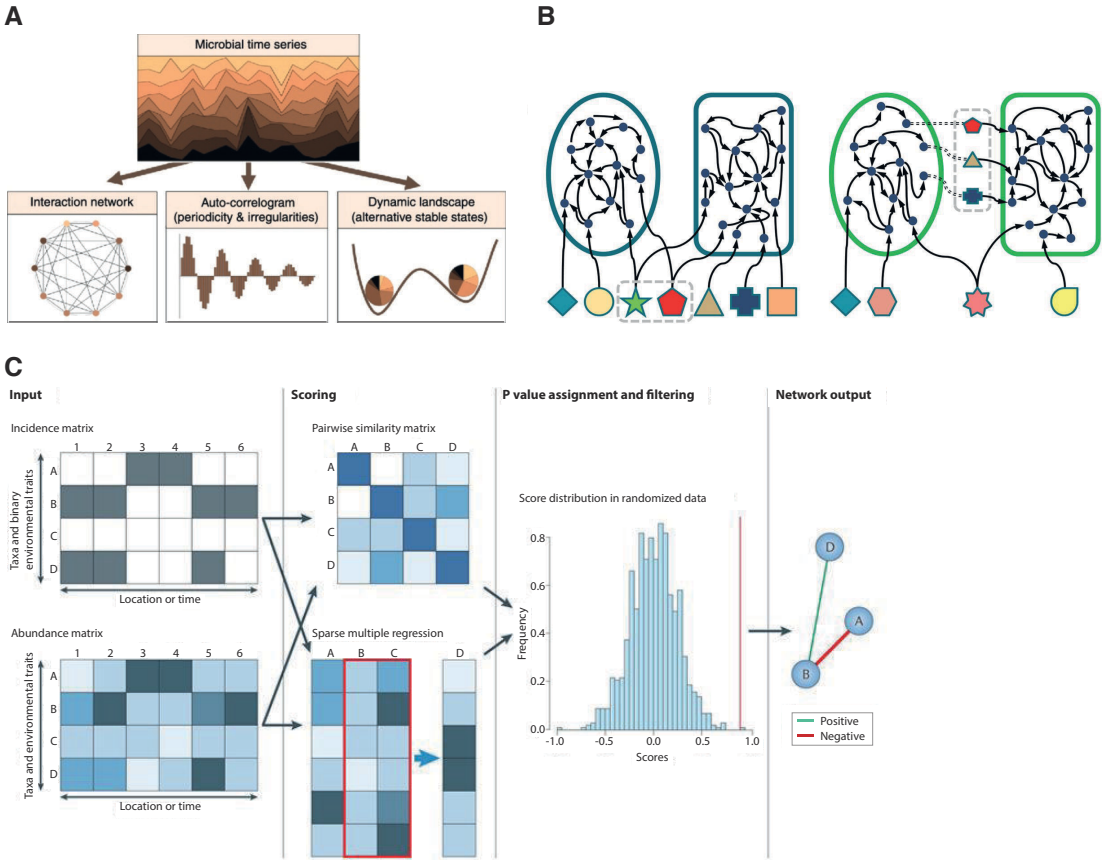


Figure 1.12: Classes of methods for the prediction of microbial interaction networks. **A** Temporal methods use longitudinal data to infer ecological relationships based on time-lagged abundance shifts. They can for instance yield insights into community dynamics such as stable states. Credit due to Faust et al. (2015). **B** Inference methods based on metabolic complementarity and redundancy compare resource requirements, inferred from genomes and metabolic models, to predict ecological dependencies. Compared to abundance-based predictions, this evidence can be more direct. Adapted with permission from Levy and Borenstein (2013). **C** Co-occurrence methods compute association scores between species based on cross-sectional data (employing a variety of strategies) and report statistically significant positive or negative associations, which can be indicators of ecological interactions. Because they only require readily available cross-sectional data, these methods are a popular choice for ecosystem analyses. Adapted with permission from Faust and Raes (2012).

also been developed (see Figure 1.12 B). The key idea of these approaches is that many microbial interactions are expected to be driven by metabolite exchange, where one species produces compounds that can be used by another (complementarity), or by similar metabolic requirements, which may lead to competition (redundancy) (Levy and Borenstein, 2012). Powered by the growing numbers of known metabolic pathways (many from novel environmental genomes, Tracanna et al. (2017)), these metabolic complementarities and redundancies can be detected across more and more species. Some of these approaches additionally allow the incorporation of extrinsic environmental factors, such as nutrients in surrounding culture media, which enables the prediction of environment-specific interactions (Levy and Borenstein, 2012).

Examples of current tools of this class are RevEcoR (Cao et al., 2016), which computes complementarity indices based on a metabolite secretion and consumption model, MMinte (Mendes-Soares et al., 2016), which utilizes flux balance analysis (FBA) to model growth effects conveyed by metabolic complementarity and redundancy, and

the previously mentioned MetaMIS method, which combines generalized Lotka-Volterra models with a metabolic complementarity index. Ecology studies employing metabolic models have for instance detected widespread metabolic complementarity across various microbial environments, which appears to provide particular advantages under oligotrophic conditions and strongly matches species co-occurrence patterns (Zelezniak et al., 2015). Metabolic methods furthermore revealed a strong role of habitat filtering relative to competition in the human gut microbiome (Levy and Borenstein, 2013). However, despite recent advances, sufficiently complete genomes are not available for the majority of microbial species and furthermore, metabolic maps are still notoriously hard to construct and thus generally incomplete (Caspi et al., 2013). Additionally, large numbers of predicted genes, in particular in novel genomes from more exotic environments, cannot be assigned any known function (Overmann et al., 2017), which precludes the application of genome-informed methods to such organisms and environments. Finally, while metabolic dependencies are important mediators for ecological interactions, other relevant factors, such as antibiotic production and mutual protection, are missed by these approaches. Similar to tools for longitudinal data, no comprehensive comparative benchmark studies have been conducted so far.

The third class of methods statistically infers microbial interactions from cross-sectional data, based on co-occurrence and co-avoidance patterns (see Figure 1.12 C), and currently constitutes the most diverse and utilized group of tools. In fact, checkerboard (presence-absence) patterns—a variant of co-occurrence—have already been used in the 1970s to infer ecological interactions between bird species across different islands (Diamond, 1975). The key assumption of co-occurrence methods is that statistically significant positive (co-occurrence) and negative (co-avoidance) relationships across large numbers of cross-sectional samples (typically sequencing samples in the context of microbial ecology) are indicative of species-species interactions (Faust and Raes, 2012). Since cross-sectional samples currently constitute by far the most abundant and heterogeneous type of data, these approaches allow predictions for a vast variety of habitats and conditions. Furthermore, they can be used infer interactions between novel and rare microbes (given proper care, see subsection 3.1.3), which may for instance lack complete genomes or informative time-series data.

While a number of earlier studies used unadjusted association measures, such as Pearson or Spearman correlation (Arumugam et al., 2011; Barberán et al., 2012; Chaffron et al., 2010), it has now been recognized that the compositional nature of microbial abundances necessitates more careful methods for association prediction (Gloor et al. (2017); see subsection 1.3.2). Examples of more compositionally-robust measures include for example the Bray-Curtis dissimilarity and the Aitchison distance (Röttjers and Faust, 2018). Alternatively, methods that try to address compositionality issues through log-ratios have been developed, such as SparCC (Friedman and Alm

(2012); or its more scalable reimplementation, fastspar (Holt et al., 2018)), SPIEC-EASI (Kurtz et al., 2015), CCLasso (Fang et al., 2015) and a recent method proposed in (Egozcue, 2018). Yet other approaches try to address sparsity and compositionality issues directly through their statistical model—a strategy employed for instance by mLDM (Yang et al., 2017), gCoda (Fang et al., 2017) and BAnOCC (Schwager et al., 2017)—or use a permutation-based re-normalization scheme (CoNet, Faust and Raes (2016)).

Further distinguishing properties between methods exist: while CoNet uses an ensemble method to exploit complementary advantages of multiple association measures, BAnOCC utilizes a Bayesian approach to provide uncertainty estimates for predicted interactions. MENA (Deng et al., 2012), a method based on Random Matrix Theory, aims at reducing arbitrary threshold requirements to increase robustness to noise. MPLasso (Lo and Marculescu, 2017), on the other hand, complements the traditional graphical Lasso approach with prior information gained from text mining of scientific literature to improve network recovery. eLSA, mentioned in the discussion of temporal methods, can also be applied to cross-sectional data and makes efficient use of technical replicates. mLDM, CoNet and a method proposed in Wadsworth et al. (2017) have been developed with non-microbial factors (e.g. physicochemical variables) in mind and allow the inclusion of these variables into predicted networks. A more recent method, BDMMA (Dai et al., 2018), additionally accounts for batch effects. Previously mentioned methods based on graphical models, i.e. SPIEC-EASI, mLDM, BAnOCC, gCoda, BDMMA and the method from Wadsworth et al. (2017), as well as another approach based on comparative steady-state analysis (Xiao et al., 2017), furthermore aim at reducing spurious associations, caused for instance by shared ecological dependencies (Röttjers and Faust (2018); see subsection 1.3.2).

Co-occurrence methods have yielded a number of intriguing insights, including for instance the detection of putative keystone species in a large number of environments (Banerjee et al., 2018), phylogenetic assortativity patterns across habitats (Chaffron et al., 2010; Faust et al., 2012; Kurtz et al., 2015), interactions associated with niche creation and succession patterns in the oral cavity (Faust et al., 2012), and structural changes in association networks of soil communities in response to shifts in carbon dioxide levels (Zhou et al., 2010).

A disadvantage of cross-sectional methods is that statistical co-occurrence and co-avoidance signals are more indirect than longitudinal or genomic information, leading to less certain edges and making the prediction of candidates for causal relationships generally harder (Faust and Raes, 2012). Furthermore, the decreased information value of cross-sectional data allows these methods to typically only predict symmetric, undirected edges. This property neglects the possibility of asymmetric ecological relationships, which are known to play important roles in plant-animal networks

(Bascompte et al., 2003; Bascompte et al., 2006). In addition, most co-occurrence methods are restricted to structural predictions, while temporal approaches also allow the inference of dynamics (Faust et al., 2015).

Several broad comparative benchmarks of cross-sectional approaches have been reported, which highlighted discrepancies between methods in terms of prediction accuracy, false discovery rate, robustness to compositionality and the prediction of general graph properties, such as hub species (Röttjers and Faust, 2018; Weiss et al., 2016). These observations emphasize the need for experimental gold-standard interactions to more properly distinguish accurate from less reliable tools (Faust and Raes, 2012).

1.3.2 Spurious associations and their sources

A problem faced by co-occurrence methods is that not all statistical associations between microbes are necessarily driven by ecological interactions, but may rather be caused by shared ecological dependencies, abiotic factors and technical artifacts. Such spurious associations can drastically inflate the density of a network, blurring it's true interaction-based topology and making ecological interpretation hard or impossible (Layeghifard et al., 2017; Röttjers and Faust, 2018).

A major reason for spurious associations is the compositionality of microbial abundance data, which pervades all types of microbial ecology analyses (Gloor et al., 2017), but in particular affects co-occurrence methods (Egozcue, 2018; Friedman and Alm, 2012). The underlying reason is that a variety of largely arbitrary technical factors, including for instance differences in extracted amounts of DNA or variations in sequencing quality, affect the total number of reads sequenced for a sample. Due to these sequencing depth-related fluctuations, counts are not directly comparable between sequencing samples (Faust and Raes, 2012). Total sample sum (TSS) normalization is commonly used to convert read counts into relative abundances (Goodrich et al., 2014; Knight et al., 2018), which explicitly reflect the compositional nature of the data (i.e. account for differences in sequencing depth). However, even after TSS normalization, the arbitrary sum constraint introduced by sequencing forces abundances into a simplex in Euclidean space (see Figure 1.13 A). Thus, changes in the count of one species will induce changes in others, even if its real absolute cell count stayed the same across samples (Gloor et al., 2017). Similar to TSS, this also affects the use of rarefaction normalization: while library size per sample is equalized, the sum constraint is still in effect (albeit normalized) (Gloor et al., 2017).

As has been noted by Karl Pearson already in 1897, traditional correlation measures are destined to produce incorrect results on such compositional data sets (Pearson, 1897). While some studies tackle this problem experimentally by estimating total cell

counts or biomass to properly calibrate read numbers (Gifford et al., 2011; Nakatsuji et al., 2013; Props et al., 2017; Vandeputte et al., 2017a), such approaches are still rarely used. The main reasons for this are likely the requirement of specific equipment, not available to all research labs, and increased expenses compared to only performing sequencing.

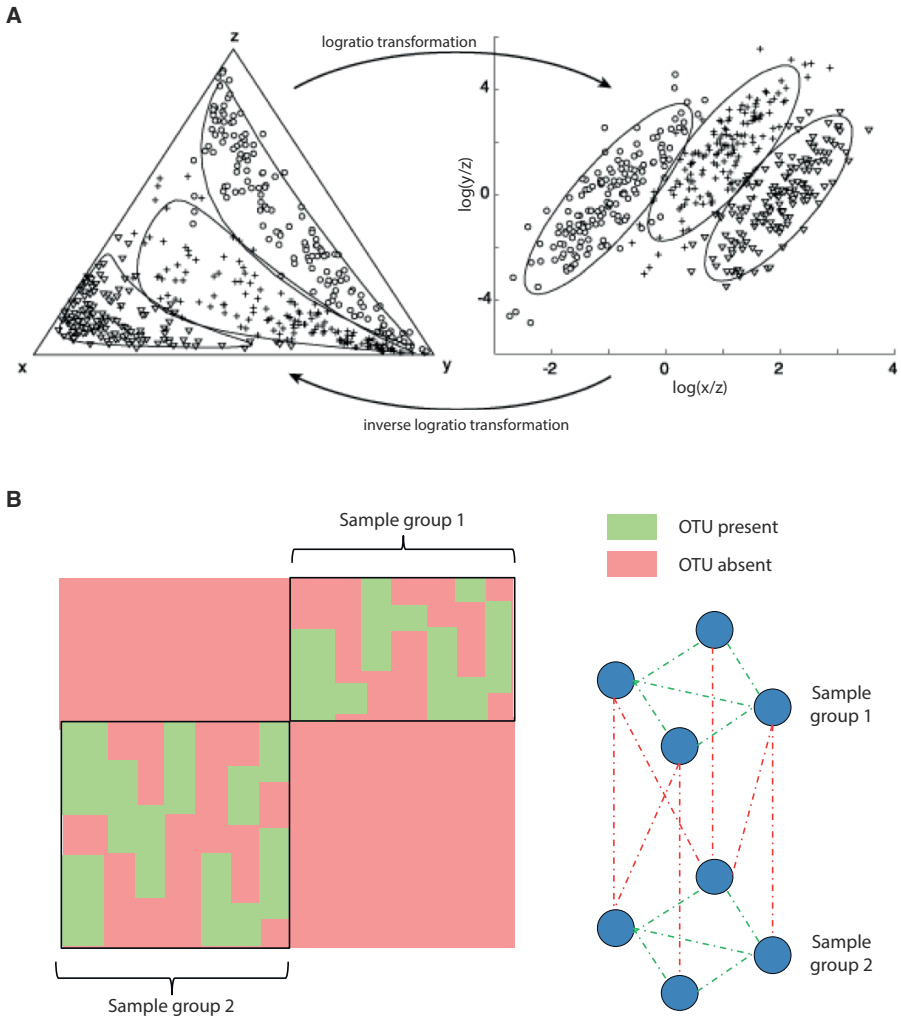


Figure 1.13: Sources of spurious associations in co-occurrence networks. A Microbial abundances are compositions and thereby constrained to the simplex. Valid statistical analysis, such as association calculations and hypothesis tests, requires transformation of the data into Euclidean space or the usage of operations defined by compositional geometry. Adapted from Vermeesch (2006). **B** Heterogeneous data is composed of systematically different groups of samples, for instance from different habitats, collected under different conditions or generated using different experimental protocols. This leads to group-dependent changes in abundances and, notably, the introduction of potentially extensive structural absences. Both have the potential to introduce substantial numbers of spurious positive and negative associations.

To facilitate the handling of compositional data, a vast body of literature has been produced in the field of compositional mathematics to make this type of data amenable to traditional Euclidean multivariate analysis, including correlation, regression and ordination techniques (Aitchison, 1982; Pawlowsky-Glahn et al., 2015). A typical approach, first introduced and championed by Aitchison in the context of Geology (Aitchison, 1981), is the use of log-ratios instead of relative abundances. Log-ratio-based techniques exploit the fact that, while relative abundances may change in

response to shifts in other compositional abundances, log-ratios will remain unaffected (Aitchison et al., 2000). This is for instance reflected in the Aitchison distance, which measures the distance between two variables in terms of variation in their log-ratio (Aitchison, 1982). In general, logarithmic transformations, such as the additive log-ratio transformation (alr), the centered log-ratio transformation (clr) and the more recently proposed isometric log-ratio transformation (ilr), can be used to transform compositional data back into Euclidean space and thus enable Euclidean multivariate analysis (Aitchison, 1982; Egozcue et al., 2003; see Figure 1.13 A).

While this transformation frees the data from the simplex, there are potential difficulties in interpreting these transformed values (Egozcue et al., 2003). Though in practice more challenging, specialized compositional geometry thus defines a number of operations that allow the sound analysis of data directly within the simplex space (Aitchison, 1982). Another potential problem of log-based approaches is that they are not directly applicable to data with zeros, because they rely on logarithmic calculations (Egozcue et al., 2003). This can create considerable problems for microbial abundance data sets since these typically feature large fractions of absences (Kaul et al., 2017).

To make such data amenable to log-ratio approaches, various schemes of data imputation have therefore been devised (Martín-Fernández and Thió-Henestrosa, 2006; Martín-Fernández et al., 2003; Tsilimigras and Fodor, 2016), most prominently different types of pseudo-counts with individual advantages and disadvantages. While this comparatively simple class of methods provides reasonable performance in practice (Friedman and Alm, 2012; Mandal et al., 2015), the choice of pseudo-counts can nonetheless affect results in microbiome studies (Costea et al., 2014; Kaul et al., 2017). Alternatively, other approaches explicitly model compositionality within the statistical frameworks they employ, but in consequence rely on various assumptions (Fang et al., 2017; Kaul et al., 2017; Schwager et al., 2017; Yang et al., 2017).

While approaches based on log-ratios have generally produced good results in simulation studies testing normalization effects on differential abundance detection and co-occurrence network inference (Friedman and Alm, 2012; Kurtz et al., 2015; Mandal et al., 2015; Weiss et al., 2016; Weiss et al., 2017), comparisons to compositionality-mitigating experimental approaches are so far lacking. For instance, Vandeputte et al. found striking differences between networks based on compositionally-calibrated and uncalibrated sequencing data but did not consider networks based on log-ratio transformed data (Vandeputte et al., 2017a).

Apart from compositionality, a variety of biological and technical factors can induce spurious associations (see Figure 1.13 B; Figure 1.15 A). The first type of confounders are shared ecological dependencies, which emerge when two microbes (A and B) do not directly interact ecologically, but are dependent on the same partner (C). C may for instance produce secondary metabolites consumed by A and B or provide extracellular

enzymes that make important elements, such as iron or nitrogen, bioavailable to both A and B (Falkowski et al., 2008; Fischbach and Sonnenburg, 2011; Griffin et al., 2004; Röttjers and Faust, 2018). Co-occurrence approaches naive to shared ecological dependencies may report the detected association signal between A and B as a possible interaction, even though it is indirect. Similarly, if C generates environmental conditions unsuitable for A and B, for instance by dropping the environmental pH through waste products (García et al., 2017) or by producing substances (e.g. antimicrobials) toxic to both A and B (Lemos et al., 1991; Long et al., 2013), this will induce a spurious positive association between A and B (in addition to the negative associations A-C and B-C, which represent genuine interactions). Similarly, if A and B are being preyed upon by the same predator C, this also induces a linked fluctuation of abundances in A and B. Independent of the actual interaction type, shared ecological dependence effects may induce particularly numerous associations between neighbors of ecological hub species, because these hubs have large numbers of interaction partners that consequently all share the same ecological dependency (i.e. the hub; Berry and Widder (2014)).

Another source of spurious interactions are shared environmental dependencies, which can—similar to shared ecological dependencies—induce indirect co-occurrence and co-avoidance patterns between microbes (Röttjers and Faust, 2018). For instance, similar requirements with regards to temperature, pH, light conditions, phosphate and iron levels may result in spurious positive associations between species, while differences in environmental preference can result in spurious co-avoidance patterns. Consequently, ecological interactions can be hard to distinguish from environmental conditions using co-occurrence alone (Pascual-García et al., 2014; Röttjers and Faust, 2018). Similarly, the presence of different habitats (or niches) within the same data set can generate positive associations between microbes specific to the same niche, and negative associations between microbes from different niches, even in the absence of any ecological interactions. The severity of spurious associations in environmentally heterogeneous data sets even lead to the suggestion that clusters in co-occurrence networks inferred from such data (using common co-occurrence-based tools) should be generally considered more as environmental niches, rather than groups of tightly interacting species, and that specific sampling strategies are necessary to circumvent niche-effects (Röttjers and Faust, 2018).

Apart from ecological and environmental confounders, experimental conditions can also induce both positive and negative spurious associations. This includes systematic differences in studied groups, for instance variations in diet or antibiotics usage in human and animal subjects (Debelius et al., 2016), but importantly also technical factors and batch effects (Dai et al., 2018; Leek et al., 2010; Salter et al., 2014; Tremblay et al., 2015). For instance, if two microbes are preferentially amplified by the same primer (i.e. are affected by the same primer bias) in a data set generated

with multiple primers, they will be indirectly positively associated. Similarly, if one species is preferentially amplified by one primer type and the second species by another (or even unaffected by primer bias), they will show a spurious co-avoidance pattern. The same principle applies to mixes of amplicon and WGS samples within the same data set: species that can only be detected with WGS will be indirectly positively associated to each other and negatively associated to correctly amplified species. Other technical factors, such as extraction methods, can result in similar problems, i.e. species preferentially extracted by one method will have positive spurious associations with each other and negative spurious associations with species showing different extraction specificity.

Sources of heterogeneity, whether due to different habitats, physicochemical factors, treatments or technical factors, are becoming more important as the number of globally distributed sequencing studies grows (see subsection 1.2.4). As described earlier in this subsection, large quantities of spurious associations introduced by heterogeneity make the detection of genuine ecological interactions hard and inflate the density of inferred co-occurrence networks, hampering interpretation of the resulting models (see Figure 1.13 B). One way to tackle this problem is a split-and-merge approach, in which separate networks per known group (typically habitats or conditions) are computed and subsequently merged into one global model (Faust and Raes, 2012; Lima-Mendez et al., 2015). While this method rigorously prevents spurious associations induced by known stratified groups, it lacks sensitivity for more subtle patterns, such as interactions of cosmopolitan species, which may only be appreciated across multiple habitats (Pascual-García et al., 2014), or rare species, which require higher statistical power for detection (Banerjee et al., 2018; Lynch and Neufeld, 2015). Additionally, despite efforts to standardize metadata in microbiome research (Yilmaz et al., 2011), annotations in public databases are still notoriously incomplete and inconsistent across studies (Lagkouvardos et al., 2016; Mitchell et al., 2018; Pasolli et al., 2017). This makes the split-and-merge approach largely inapplicable to the wealth of public sequencing data since group-membership cannot be reliably inferred. Notably, even within a single study that is carefully stratified by known habitats and conditions, unmeasured confounders (latent variables) can introduce heterogeneity in the form of unlabeled sub-niches, not obvious to the investigator and thereby not accounted for in the merge-and-split approach. Finally, shared ecological dependencies are not addressed by this method, because these are genuine biological phenomena and thus create spurious associations also within homogeneous data sets.

Nonetheless, combining diverse studies into aggregated data sets promises the opportunity to detect emergent large-scale, cross-environmental trends (e.g. global topological motifs) that would not be noticeable within single studies and may reveal subtle patterns involving cosmopolitan and rare species. Hence, the development of

principled methods for tackling spurious associations within large-scale, cross-study data sets is paramount.

1.3.3 Probabilistic Graphical Models as a framework for ecological network inference

In the network structure learning community, it has long been appreciated that univariate association networks include both direct and indirect associations: the former driven by direct, mechanistic relationships between variables, while the latter are induced by indirect effects, i.e. shared dependencies on the same variable. Deeper characterization of this insight led to the mathematical formalization of the classic statistical mantra that correlation does not necessarily imply causation (Buntine, 1996; Heckerman, 1990; Pearl, 1988; Spirtes et al., 2000). This process has sparked the development of a vast variety of dedicated methods to specifically distinguish direct from indirect associations. These methods thus aim to predict more sparse and interpretable network models that are closer approximations to the real mechanistic relationships of the underlying systems.

Currently, the most widely used class of such methods are structure learning techniques for Probabilistic Graphical Models (PGMs, see Koller et al. (2009) for a detailed discussion). Methods of this type have been highly successful at tackling data mining and network prediction problems across diverse domains, such as image and speech recognition, medical diagnosis, natural language processing and fault diagnosis (see Koller et al. (2009) and citations therein). Notably, PGM approaches were furthermore effectively used to predict the structure of biological networks, such as gene regulatory networks (Friedman et al., 2000; Narendra et al., 2011), and were more recently highlighted for their potential in microbial interaction network reconstruction (Layeghifard et al., 2017).

PGM methods exploit the concept of conditional independence, which holds if two univariately associated variables are rendered statistically independent given a set of other variables with explanatory power for this relationship (Bühlmann et al., 2014; Pearl, 2010). For instance, if two species (A and B) are indirectly associated due to a shared ecological dependence on a third species (C), the univariate association between A and B will disappear when conditioned on C (in statistical terms: species A and B are conditionally independent given C). In contrast, if the association between A and B persists after conditioning on C, the relationship is considered to be direct. This concept also applies to mixes of species and abiotic variables: if two species are indirectly associated, for instance due to their shared environmental preference for high temperature or a specific habitat, but are not ecologically interacting, conditioning on the relevant abiotic variables can identify the observed association as indirect.

Also, technical factors and treatment conditions can be accounted for in this way—as long as the variables are measured and included into conditional independence tests, indirect associations can generally be explained away. A limitation for these methods is, however, that they generally necessitate higher data quantities to reliably determine whether conditional dependence holds (Schlüter, 2012; Spirtes et al., 2000). While this may hamper their applicability to single studies, the staggering size of modern cross-study microbiome data sets (see section 1.2.4) makes statistical power less of an issue.

The most prominent classes of PGMs are Markov Random Fields (MRFs) and Bayesian Networks (BNs), which are undirected and directed graphs, respectively. Either class features a number of dedicated structure learning algorithms and the choice of PGM type depends on the nature of the problem (Koller et al., 2009), i.e. whether it can reasonably be modeled using a directed graph, implying causal relationships between variables (see Figure 1.14 A), or not. While both MRFs and BNs have been used in a variety of different fields, including the modeling of ecological communities (Clark et al., 2018; Milns et al., 2010), the directionality implied by BNs is arguably preferable for ecological networks since asymmetry is an important property of ecological relationships[23] (Bascompte et al., 2003; Bascompte et al., 2006). Furthermore, BN learning methods excel at the prediction of causal links (Heckerman, 1997; Spirtes et al., 2000), which are of key interest in research settings (e.g. ecology), but are typically not a focus in engineering applications, where MRFs shine (Koller et al., 2009).

While causal structure can be predicted through interventions or randomization trials, obtaining experimental data at scales sufficient for reliable causation analysis in large systems can be prohibitively laborious or ethically questionable (e.g. forcing randomly selected people to smoke, Spirtes et al. (2000)). Fortunately, given several assumptions, BN structure learning algorithms can allow the inference of causal relationships from observational data alone, in the sense of an operational definition of causality that nonetheless closely matches the idea of Randomized Controlled Experiments (Aliferis et al., 2010a; Spirtes et al., 2000). In particular, these methods require i) the variables in the network to be causally sufficient, i.e. the absence of unmeasured variables that directly causally influence more than one variable in the system, and ii) the faithfulness assumption, i.e. the conditional independence structure of the joint probability distribution of the variables must be perfectly representable by a DAG (Spirtes, 2010).

Given these sufficient conditions, a variety of methods can be employed to learn

[23] However, in order to allow for reliable and efficient causal inference, BNs require the underlying graph to be acyclic (i.e. a directed acyclic graph or DAG). Thus, only causal asymmetries can be modeled with BNs, whereas bi-directional asymmetries, in which each direction has a different sign or weight, cannot be described.

causal graph structure directly from observational data (described in detail in Koller et al. (2009) and Spirtes et al. (2000)). A widely-used class are search-and-score approaches, which scan the space of feasible DAGs for graphs that optimize a scoring criterion, typically high marginal or penalized likelihoods for generating the input data. Popular scores are the Minimum Description Length (MDL), Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), which all penalize complex models and thus implement an Occam's razor approach: make the model as complex as necessary, but not more complex. For the search step, various optimization algorithms can be used, such as greedy hill climbing, tabu search or basin flooding, which may start from either a random or informative initial structure. As an alternative to search-and-score algorithms, methods based on Markov chain Monte Carlo (MCMC) can be utilized, which draw samples from the posterior joint distribution of all possible graph structures (conditioned on the training data) to approximate the true network.

Advantages of search-and-score and MCMC methods include that they result in fully directed graphs and learn the parameters of the underlying joint probability distribution in conjunction with graph structure. They thus result in fully specified models that can be used to make predictions (Figure 1.14 A). However, the data can typically be described similarly well by a large number of structures (or even equally well, i.e. if the models belong to the same equivalence class, Koller et al. (2009)). In order to avoid biased or imprecise structure predictions and to furthermore estimate uncertainties given the data, large amounts of search space exploration or MCMC sampling are thus necessary (Koller et al., 2009). Furthermore, finding accurate structures with these methods has in general a runtime complexity that is super-polynomial in the number of variables (is NP-Hard), rendering it feasible only for comparatively small systems (Chickering et al., 2004; Spirtes, 2010).

In order to reduce computational complexity, another class of structure learning methods has been proposed: the constraint-based approaches. These algorithms aim at iteratively detecting local conditional independence relationships between variables, i.e. the set of directly associated neighbors or Markov blanket (MB) of each variable (see Figure 1.14 B), which via the Markov condition[24] correspond to the structure of the underlying graph (Spirtes et al., 2000). Conditional independencies are detected through the sequential application of statistical tests, commonly based on partial correlations with the Fisher z-transformation (for continuous variables) and three-way χ^2 or mutual information tests (for discrete variables) (Aliferis et al., 2010a).

The PC algorithm, a prototypical constraint-based method, proceeds by starting with the fully connected graph and subsequently testing for each edge, whether the connecting nodes can be rendered conditionally independent by conditioning on any

[24] In brief, the Markov condition states that variables must be independent of all other non-parent and non-descendant variables within the graph, conditioned on the set of parents and descendants.

combination of local neighbors (Spirtes et al., 2000). After this step, only edges that could not be rendered conditionally independent remain, which thus represent the set of predicted direct associations (called the undirected "skeleton" of the BN). In order to infer partial directionality of the edges (i.e. a partial DAG or PDAG, corresponding to an equivalence class of BNs), a set edge-directing rules can subsequently be applied to these undirected links (Spirtes et al., 2000).

The advantage of constraint-based algorithms is that, because they solve the conceptually simpler problem of partially directed structure prediction (also disregarding distribution parameters), they can learn the structure more reliably and computationally efficiently than search-and-score or MCMC algorithms. In order to convert the PDAG predicted by constraint-based approaches into a full DAG, and also learn the full parameters of the BN, the PDAG structure can subsequently be used to constrain the parameter space for search-and-score algorithms, as realized for instance in Greedy Equivalence Search (GES, Chickering (2002)). Such combined approaches ("hybrid" algorithms) can make model inference much more tractable, while also increasing the quality of the learned network compared to vanilla search-and-score methods (Tsamardinos et al., 2006). Alternatively, parameters of the full joint probability distribution can be learned for complete DAGs through efficient factorization approaches, which exploit the local graph structure to simplify marginalization computations (Koller et al., 2009).

While the PC algorithm is conceptually simple, it requires excessive amounts of statistical tests, since all edges have to be tested conditioned on all combinations of neighbors to produce the final network. However, under broad assumptions (Aliferis et al., 2010a), many tests are likely to be redundant and may safely be skipped, a realization that lead to the development of a multitude of heuristics, such as the GS, IAMB, K2MB, MMPC and (si-)HITON-PC algorithms (summarized in Aliferis et al. (2010a)). Some of these heuristics, for instance MMPC and si-HITON-PC, can infer direct edges within very large systems consisting of many thousands of variables (assuming sparseness and bounded maximum node degree), while providing guarantees with respect to correctness and soundness in the sample limit. In particular, just as the PC algorithm, they guarantee the asymptotic correctness of the learned structure (i.e., that it is statistically indistinguishable from the true structure) under causal sufficiency and faithfulness assumptions (see above), as well as given reliable statistical tests[25] (Aliferis et al., 2010a).

Successful applications of these algorithms include the prediction of large networks from diverse domains, such as drug-drug interactions (Duda et al., 2005), cancer diagnosis (Sboner and Aliferis, 2005) and gene regulation (Narendra et al., 2011).

[25]Notably, some heuristics (for instance algorithms based on GLL) tend to predict largely correct structures even if these assumptions are not perfectly met (Aliferis et al., 2010a; Aliferis et al., 2010b).

Furthermore, in a variety of benchmarks, they were found to predict structures of similar or better quality compared to non-heuristic approaches, at substantially reduced runtimes (Aliferis et al., 2003; Tsamardinos et al., 2003a; Tsamardinos et al., 2003b; Tsamardinos et al., 2006).

Generalized Local Learning (GLL, Aliferis et al. (2010a) and Aliferis et al. (2010b)) has been proposed as a principled, general framework for the systematic construction of such heuristic MB induction algorithms. GLL instantiations for particular choices of inclusion heuristic, elimination strategy and interleaving strategy yield for instance the previously mentioned MMPC or si-HITON-PC algorithms. Important in the context of this thesis, GLL yields local, causal feature selection algorithms: under previously mentioned assumptions, the identified MB of a variable corresponds to its causally interpretable neighborhood of predictive variables (see Figure 1.14 B). This principled strategy for feature selection showed advantages in diverse prediction tasks, where feature sets predicted by GLL methods achieved the highest causal interpretability and parsimony (i.e. the best trade-off between predictiveness and number of features) on synthetic data across a variety of systems and variables (Aliferis et al., 2010a; Aliferis et al., 2010b). In contrast, feature sets predicted by non-causal algorithms, such as univariate filtering methods and feature selection wrappers (e.g. Recursive Feature Elimination (RFE)), the latter utilizing widely-used machine learning classifiers such as SVMs (Support Vector Machines) and Decision Trees for feature selection, showed increased feature redundancy, leading increased numbers of weakly relevant or irrelevant features (in the causal sense, Aliferis et al. (2010a) and Aliferis et al. (2010b); see Figure 1.14 B).

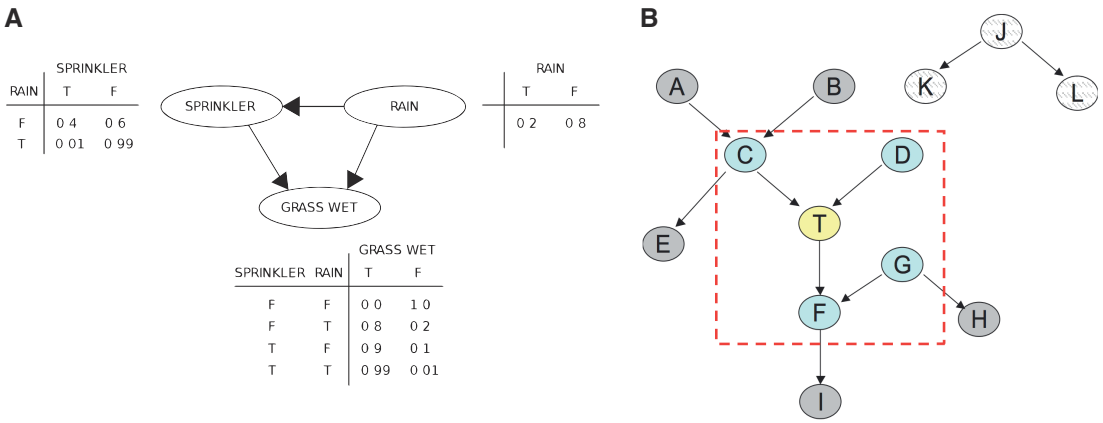


Figure 1.14: Bayesian networks and the GLL-LGL framework. A A simple example of a Bayesian network. Observing that the grass is wet allows inferring the probability that it rained recently, also accounting for the chance of the sprinkler causing the wetness instead of rain. Notably, the sprinkler's activation probability depends on whether it rained or not. Individual conditional probabilities of the joint probability distribution underlying the graphical model (i.e. its parameters) are provided through factor tables as shown here. Credit due to Wikimedia Commons (2006). **B** Important concepts for the GLL-LGL framework by Aliferis et al (Aliferis et al., 2010a; Aliferis et al., 2010b). For a target variable T (yellow), GLL aims to identify its set of directly associated, relevant predictors (i.e. its causally optimal feature set (or MB), displayed in blue). Weakly relevant predictors (grey) are only indirectly associated with T and redundant since they carry less direct information than the Markov Blanket. Irrelevant predictors (white) provide no information on the state of T. In LGL algorithms, local MBs are computed for all variables by independent application of a GLL algorithm of choice, followed by combining individual neighborhoods into a global network. Credit due to Aliferis et al. (2010a).

Local neighborhoods predicted by GLL can subsequently be combined using a global rule, such as the OR or AND rule, to connect two nodes if i) either or ii) both are found within each other's neighborhoods, which extends GLL to local-to-global learning (LGL, Aliferis et al. (2010b)). The GLL-LGL framework thus yields both local feature selection and global network prediction algorithms, making it useful for the detection of locally optimal predictive features, but also for the inference of sparse and resolved global network structure. The framework is furthermore highly flexible, allowing a number of computational extensions and optimizations, such as improved heuristics (adjusting the trade-off between correctness and computational tractability), optimized statistical tests, efficient caching schemes and parallelization (Aliferis et al., 2010b).

Recently, several PGM methods have been proposed for predicting direct associations in microbial co-occurrence networks (see subsection 1.3.1). Typically, these tools learn the precision matrix of interactions in a MRF model either through different Lasso-based approaches (such as the graphical Lasso or the Meinhausen and Bühlmann algorithm (Kurtz et al., 2015)), which employ sparse regularization to account for deficient rank in the input OTU matrix, or alternatively by MCMC sampling. While MCMC-based approaches are notoriously expensive (Andrieu et al., 2003; Sharma, 2017), Lasso-based methods require either optimization-based search through a potentially extensive parameter space (quadratic in the number of OTUs) or rely on solving high numbers of regularized regression problems (equal to the number of OTUs). In addition, Lasso-based strategies have to be repeatedly applied in a cross-validation setup to identify optimal regularization terms, and in some approaches (e.g. SPIEC-EASI), furthermore across many sub-samplings of the input data (bootstrapping).

In contrast, GLL-derived algorithms only consider a subset of heuristically determined candidates for incorporation into a neighborhood and, through incremental inclusion decisions based on lower-order statistical tests[26], identify directly associated neighbors only among those candidates. While the number of required tests can still be substantial (exponential in the number of direct neighbors of a variable), general sparsity assumptions—also made by SPIEC-EASI and SparCC, but for other reasons—allow these heuristic algorithms to infer graph structure at substantially increased speeds. As mentioned previously, MRF-based approaches furthermore do not incorporate the concept of directionality and hence do not attempt to predict causal relationships, in contrast to BN-based methods (such as GLL-LGL instantiations). For further discussion of differences between MRF- and constraint-based BN methods in the context of microbial co-occurrence networks, see Layeghifard et al. (2017).

[26] which also address dimensionality issues

1.4 Research goals

"All we have to decide is what to do with the time that is given us."

—Gandalf (*Lord of the Rings*[27])

1.4.1 General aims

The primary research goal of this thesis is the development of a new, co-occurrence-based method to predict interpretable and comprehensive microbial interaction networks, across environments and conditions, with explicit inclusion of non-microbial factors. Here, interpretability means that predicted networks should be largely free of spurious associations, as caused for instance by shared ecological and environmental dependencies or technical confounders (see subsection 1.3.2 for a detailed discussion). Comprehensiveness, on the other hand, describes the notion that networks should not be restricted to interactions of highly abundant and prevalent species within a single homogeneous environment, but instead also integrate rare and cosmopolitan species across heterogeneous studies. Rare species may play important roles in ecosystem function, but are usually removed in network prediction analyses due to issues related to statistical power and detection limits, while interactions of cosmopolitan groups (i.e. generalist species) may only become detectable in their entirety within cross-habitat data sets. Being comprehensive furthermore implies the inclusion of more subtle interactions between abundant specialist species, which may only become apparent under certain experimental conditions (e.g. specific host diets and medications), as well as interactions that may only be detected by particular technologies, such as specific primers or WGS sequencing. Finally, comprehensiveness also involves the explicit inclusion of important non-microbial factors, for instance physicochemical variables (e.g. pH, temperature), habitat types (e.g. marine, soil, human gut) and experimental conditions, to enable a more complete interpretation of underlying ecosystems.

If interpretable associations (in the above sense) between individual microbial species and interesting non-microbial factors (e.g. disease status) can be predicted, this naturally yields a promising method for parsimonious, i.e. non-redundant and predictive, microbial biomarker discovery. A complementary, secondary research goal of this thesis is thereby to explore a method for ecologically informed biomarker discovery, taking advantage of the improved interpretability achieved within the scope of goal one.

[27] Written by *John Ronald Reuel Tolkien*

1.4.2 Current challenges

A first problem for the prediction of comprehensive and interpretable networks (goal one) is the computational burden associated with the consistent processing large numbers of sequencing studies, crucial in particular for the "comprehensiveness" requirement in the previously discussed sense. Since different studies utilize a range of different sequencing protocols, in particular amplicon primers targeting different hypervariable regions, the use of closed-reference OTU mapping approaches is a necessary precondition for making samples comparable[28]. However, current closed-reference OTU mapping tools are not able to efficiently process samples at modern scales (> 1 million publicly available sequencing samples, see subsection 1.2.4) with reasonable time and resource investment.

Secondly, as outlined in subsection 1.3.3, interpretability issues brought about by indirect associations can be tackled in a principled way through PGM structure learning approaches. This idea is increasingly realized by the microbial ecology community and led to the recent development of PGM-based tools for co-occurrence network prediction to improve the interpretability of inferred networks. However, current methods were primarily developed for small-scale microbial ecosystems and thus do not scale to the cross-study data sets required for the prediction of comprehensive networks. Furthermore, these tools generally do not account for abiotic and experimental factors, thereby leaving spurious associations introduced by these confounders largely unchecked. Abiotic and experimental factors are of particular importance in cross-sectional studies, which can therefore not be coherently analyzed by current methods. Additionally, incomplete or inconsistent metadata annotations of publicly available microbial sequencing samples[29] can result in large numbers of false positive associations for current PGM approaches, further hampering their applicability to cross-study data sets.

With regard to microbial biomarker discovery (goal two), the main challenge is that current methods also report spurious associations between microbial species and variables of interest, driven for instance by ecological dependencies between microbes. Ideally, these indirect associations should be removed if parsimonious biomarkers with a more mechanistic interpretation are desired, as for instance when investigating candidates for disease-causing microbial species.

[28] other phylotypes, such as ESVs and *de novo* OTUs, do not satisfy this comparability requirement (see subsection 1.2.3)

[29] and also unmeasured confounders within single, well-annotated studies

1.4.3 Proposed solutions

Mapping OTUs in a scalable and consistent way is a currently difficult, but nonetheless critical step towards obtaining data sets from which comprehensive networks can be predicted. To tackle this challenge, I assisted in developing and testing a novel OTU mapping tool: MAPseq (see Appendix A). It utilizes an optimized pre-clustering approach in conjunction with efficient k-mer hashing, followed by exhaustive within-cluster search for optimal hits, to achieve markedly improved runtimes. Combined with a highly optimized implementation, this workflow allows the rapid mapping of sequences to databases with a high degree of redundancy, as typical for full-length 16S rRNA reference databases. MAPseq allowed us to analyze more than a million sequencing samples, mapped to a comprehensive full-length 16S rRNA reference database, which resulted in comparable abundance profiles for more than ten thousand highly heterogeneous microbial ecology studies (compiled into the MAPdb database). An earlier version of this database, featuring more than five hundred thousand samples, was used for the co-occurrence network analysis reported in this thesis.

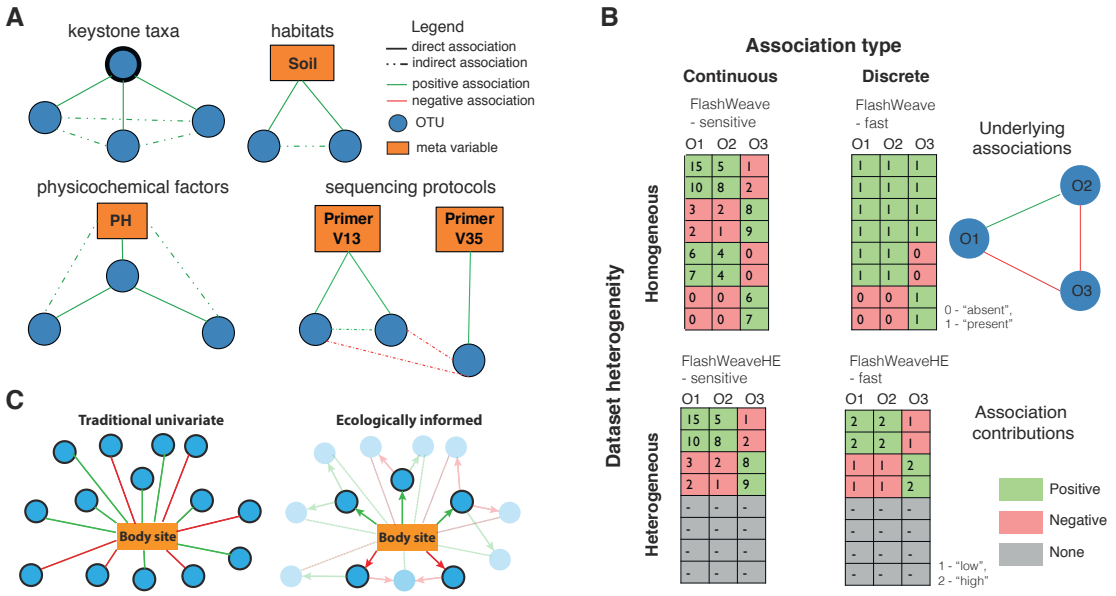


Figure 1.15: Proposed methods for inferring interpretable ecological networks and biomarkers from heterogeneous data sets. **A** Spurious associations introduced by shared ecological, environmental or technical dependencies can be tackled by a conditional independence approach, as implemented in FlashWeave. In this framework, only links that are still significant after conditioning on potentially confounding variables (e.g. physicochemical or technical factors) are reported as candidates for ecological interactions. **B** In order to also mitigate the effect of unmeasured confounders in heterogeneous data sets, FlashWeaveHE furthermore discards absence information in its statistical tests and thus removes the substantial effect of structural and sampling zeros. **C** Ecologically informed biomarkers, which can also be learned with FlashWeave, are directly associated to a variable of interest (here a human body site) and are thus more parsimonious compared to widely reported univariate biomarkers. In particular, they are depleted of indirect associations that may arise for instance through microbe-microbe interactions.

In order to tackle the remaining challenges, we developed a novel software framework for the inference of highly-resolved microbial co-occurrence networks, as well as parsimonious and interpretable biomarker discovery, from large-scale sequencing data sets: FlashWeave (see Manuscript 2.1). FlashWeave is a thoroughly

optimized implementation of the scalable GLL-LGL framework for causal inference, as proposed by Aliferis et al. (see Aliferis et al. (2010a) and Aliferis et al. (2010b); subsection 1.3.3), complemented by several critical extensions. FlashWeave allows the prediction of direct association networks, depleted for spurious associations, from data sets with hundreds of thousands of samples and tens of thousands of OTUs.

This computational efficiency was achieved through multiple innovations. Firstly, in addition to the careful and optimized implementation of a comprehensive set of heuristics and shortcuts proposed in the literature, we devised multiple novel heuristics. These computational shortcuts can profoundly speed up inference for networks with ecosystem-typical properties, such as right-tail heavy node degree distributions (see subsection 1.1.2). Secondly, we devised a novel pseudo-count scheme (adaptive pseudo-counts), which drastically reduces spurious associations that may arise during log-ratio transformation if conventional pseudo-counts are applied to rare species. Thirdly, FlashWeave introduces novel methods for handling data heterogeneity, both due to measured and unmeasured confounders, which further reduce spurious edges and enable tremendous speed-ups compared to vanilla GLL-LGL approaches.

Additionally, the flexibility of the GLL-LGL framework allows FlashWeave to seamlessly incorporate non-microbial meta variables. On the one hand, this enables the removal of spurious associations driven by these factors, but on the other hand, it also allows insights into which microbial taxa are directly affected by these variables. While this explicit modeling of abiotic factors can be tremendously helpful in microbial ecosystem interpretation, the reverse view, i.e. which microbial species are direct biomarkers for non-microbial variables of interest (modulo microbe-microbe interactions), can also reveal important patterns. This alternative, biomarker-centric view benefits from a number of desirable properties of GLL-based local learning algorithms (see subsection 1.3.3) and has been explored in detail by us in the context of predicting human body source origin from microbial content (see Manuscript 2.2).

CHAPTER 2

Manuscripts

2.1 Rapid inference of direct interactions in large-scale ecological networks from heterogeneous microbial sequencing data

Contribution: Janko Tackmann (JT) conceptualized the idea for this project, aided by all co-authors. JT furthermore developed and tested the proposed software method (FlashWeave), suggested the majority of algorithmic innovations for this method and made major contributions to the interpretation of results. In addition, JT created the visualizations found in the manuscript, wrote the initial manuscript and contributed to reviewing and editing of the final manuscript.

Rapid inference of direct interactions in large-scale ecological networks from heterogeneous microbial sequencing data

Janko Tackmann[1], João Frederico Matias Rodrigues[1], Christian von Mering[1*]

[1] Institute of Molecular Life Sciences and Swiss Institute of Bioinformatics, University of Zurich, Switzerland
*To whom correspondence should be addressed (mering@imls.uzh.ch)

Keywords: metagenomics; microbial interactions; conditional independence; co-occurrence; environmental factors; technical factors; structural zeros

Summary

The recent explosion of metagenomic sequencing data opens the door towards modeling of microbial ecosystems in unprecedented detail. Statistical prediction of ecological interactions in particular could strongly benefit from this development, however current methods are hampered by insufficient statistical resolution, limited computational scalability or by not accounting for metadata. Here we present FlashWeave, a new approach based on a flexible Probabilistic Graphical Model framework that infers highly resolved direct microbial interactions from massive heterogeneous microbial abundance data sets, with seamless integration of metadata. On a variety of benchmarks, FlashWeave outperforms state-of-the-art methods by several orders of magnitude in terms of runtime, with overall increased accuracy. We use FlashWeave to rapidly analyze a cross-study data set of 69'818 publicly available human gut samples, resulting in the largest and most diverse network of gastrointestinal microbial interactions to date. By discarding 96% of edges as indirect, FlashWeave reveals distinct patterns of biological interest, highlighting advantages of direct interactions in cross-study settings.

Introduction

Microorganisms shape virtually every aspect of Earth's biosphere. Besides their critical role in global geochemical cycles (Falkowski et al. 2008) and widespread symbiosis with all major branches of life (Oh et al. 2009; McFall-Ngai 2014; Kawaguchi & Minamisawa 2010), the tight coupling between the microbiome and human health is rapidly gaining appreciation (Carabotti et al. 2015; Thaiss et al. 2016). While the structure of microbial ecosystems is influenced by environmental factors and hosts (Bonder et al. 2016; Dyhrman et al. 2007; Krause et al. 2012), another important driving force are ecological interactions between microbes (Faust & Raes 2012; Xavier 2011), such as competition, symbiosis, commensalism, and antagonism.

The inability to (co-)culture the majority of microorganisms in the lab (Solden et al. 2016; Goers et al. 2014) makes computational tools instrumental for the prediction of ecological dependencies between microbes. Common to these approaches is the utilization of cross-sectional (co-occurrence and co-abundance (Chaffron et al. 2010; Friedman & Alm 2012; Kurtz et al. 2015)) and temporal (Stein et al. 2013; Xia et al. 2011) statistical patterns, or alternatively metabolic complementarity (Zelezniak et al. 2015; Levy et al. 2015), to infer ecological associations and construct interaction networks. Currently widespread methods are restricted to predicting pairwise interactions through univariate statistical associations (Friedman & Alm 2012; Faust & Raes 2016; Xia et al. 2011), but more recent approaches based on Probabilistic Graphical Models (PGMs) consider the conditional dependency structure between microbes to distinguish between direct and indirect interactions (Kurtz et al. 2015; Yang et al. 2017; Röttjers & Faust 2018). While PGM approaches can result in more sparse and interpretable networks, typical drawbacks include the requirement of larger data sets with sufficient statistical power and increased computational complexity. Hundreds of thousands of microbial sequencing samples from various environments around the globe are now available (Mitchell et al. 2018), alleviating the lack of statistical power,

yet this wealth of data can currently not be utilized by state-of-the-art PGM methods due to insufficient computational scalability. Furthermore, sample heterogeneity of these cross-study data sets, such as variation in habitats, measurement conditions and sequencing technology, can lead to pronounced confounding associations, typically not addressed by current methods (Röttjers & Faust 2018).

Here, we present FlashWeave, a novel approach for inferring high-resolution interaction networks from large and heterogeneous collections of microbial sequencing samples based on co-occurrence or co-abundance. FlashWeave is highly optimized for computational speed and mitigates a number of known artifacts common in cross-study sequencing data analysis, such as compositionality effects, bystander effects, shared-niche biases and sequencing biases. It furthermore allows the seamless integration of environmental factors, such as temperature and pH, to estimate their influence on studied ecosystems and to remove indirect interactions driven by them.

We compared FlashWeave to a variety of state-of-the-art methods on a wide collection of synthetic and biological benchmarks and showed that it markedly outperforms other methods in terms of speed. In addition, it achieved overall increased accuracy, in particular on heterogeneous cross-habitat data sets with large fractions of structural zeros (non-random absences driven by environmental or technical factors). We furthermore illustrated the usefulness of integrating non-microbial factors into network analysis by including habitat and primer variables into the inference of an interaction network based on the Human Microbiome Project. Finally, we applied FlashWeave to a global collection of 69'818 publicly available microbial sequencing samples of the human gastrointestinal tract, covering 488 projects. The resulting interaction network represents the most comprehensive model of ecological dependencies of the human gut to date and was inferred using minimal computational resources and time. We analyzed this network in-depth to demonstrate its consistency with previously described ecological patterns. The network furthermore unveiled candidates for uncharacterized hub OTUs and yielded a striking signal for phylogenetic assortativity, a possible indicator for pronounced kin selection.

Results

A fast and compositionally robust method for ecological network inference, capable of handling heterogeneous data

FlashWeave is based on the local-to-global learning (LGL) approach proposed by Aliferis et al. (Aliferis et al. 2010a), a constraint-based causal inference framework for the prediction of direct relationships between variables in large systems. Algorithms of this family infer the Markov blanket of each target variable T, which constitutes the directly associated neighborhood $MB(T)$ that renders all remaining variables S probabilistically independent of T. It thus removes indirect, i.e. purely correlational, associations commonly reported by wide-spread univariate methods. Related algorithms have been successfully applied in a wide range of fields, including cancer

diagnosis (Sboner & Aliferis 2005), drug-drug interactions (Duda et al. 2005) and gene regulatory network inference (Narendra et al. 2011).
FlashWeave is a highly optimized implementation of the semi-interleaved HITON-PC (Aliferis et al. 2010a) instantiation of LGL (Fig. 1 A), critically extended with several high-performance heuristics, as well as novel methods addressing data heterogeneity and state-of-the-art compositionality correction (see Text S1). The latter is essential since abundances from sequencing data constitute compositions, constrained to the simplex, which has long been known to induce artificial correlations (Pearson 1896; Aitchison 1981; Vandeputte et al. 2017).
In contrast to most other methods, FlashWeave can furthermore utilize meta variable (MV) information (Fig. 1 B), such as subject lifestyle factors, physicochemical measurements or sequencing protocol information, to further reduce spurious associations and additionally report direct relationships between Operational Taxonomic Units (OTUs) and MVs.

Increased prediction performance on a variety of synthetic data sets

Since experimentally verified biological interactions between microbes are scarcely available, we initially employed previously published frameworks that generate synthetic data with ecological structure. We compared the quality of networks inferred by two different operating modes of FlashWeave - "sensitive" (-S) and "fast" (-F) (Fig. 1 C) - to three widely used univariate inference methods (SparCC (Friedman & Alm 2012), eLSA (Xia et al. 2011) and CoNet (Faust & Raes 2016)) and three conditional methods (mLDM (Yang et al. 2017) and SpiecEasi (Kurtz et al. 2015) with neighborhood selection (MB) and inverse covariance selection (GL)).
The first group of benchmark data sets was generated with a method based on the Normal to Anything (NorTA) approach (Kurtz et al. 2015), which uses real abundance data from sequencing experiments and a custom interaction network as inputs. Synthetic OTU abundances are drawn from a target distribution fitted to the experimental data, while respecting the partial correlations provided by the input network.
The prediction quality of all methods was evaluated on such synthetic data sets with increasing numbers of samples, fitted to data from the American Gut Project (McDonald et al. 2018). Overall, FlashWeave most accurately reconstructed the input networks as measured by F1 scores of predicted edges (Fig. 2 C): across topologies, FlashWeave-S achieved a mean F1 score of 0.68, while non-FlashWeave methods ranged from 0.07 (eLSA) to 0.65 (SpiecEasi-MB), resulting in fractions between 10% and 96% compared to FlashWeave-S (mean 56%). FlashWeave-F was generally less predictive than FlashWeave-S (mean F1 score 0.62, mean fraction 62%).
In a second accuracy benchmark ("Ecological Models") we used methods presented in (Weiss et al. 2016) to generate abundance tables with a wide range of linear ecological relationships between OTUs, featuring varying degrees of sparsity and compositionality. Across all data set sizes, eLSA achieved the highest F1 scores (mean 0.76), followed by FlashWeave-S (mean 0.68, 90% of eLSA; Fig. 2 C). Notably, FlashWeave-S scores were almost identical to eLSA at the highest number of samples (mean F1 score difference < 1%). FlashWeave-S and FlashWeave-F showed comparable results (difference < 3%), while all other methods achieved mean F1 scores of 2% (SparCC) to 74% (SpiecEasi-MB) relative to FlashWeave-S (mean 62%). In both the NorTA and the Ecological Models benchmarks, FlashWeave predictions generally improved noticeably with higher sample numbers (up to 141%), indicating efficient usage of additional data.

FlashWeaveHE, which specializes on the analysis of heterogeneous data (Fig. 1 C), was compared to other methods on simulated benchmark data with increased habitat heterogeneity. To this end, we treated the three differently-sized data sets for each ecological scenario from the Ecological Models benchmark as disjoint habitats with no OTU overlaps and aggregated them into a single data set per ecological scenario (see Methods).
FlashWeaveHE-S achieved the highest F1 scores on this benchmark (mean 0.78; Fig. 3 D), followed by FlashWeaveHE-F with 0.6 and FlashWeave-F with 0.43. The best non-FlashWeave method, SpiecEasi-GL, achieved a mean of 0.25, 68% less than FlashWeaveHE-S. Notably, FlashWeaveHE modes displayed almost perfect precision (0.99) while non-FlashWeave methods ranged from 0.0007 (SparCC) to 0.2 (SpiecEasi-GL).

Improved reconstruction of literature interactions in TARA Oceans

In a study of planktonic associations in the TARA Oceans project, the authors presented a list of genus-level interactions described in the literature (Lima-Mendez et al. 2015). This set provides a gold-standard on which network inference tools can be tested, but is limited to a small fraction of the total marine micro-eukaryotic diversity and likely incomplete. It thus can only be used to benchmark recall on a restricted subset of true positive interactions, but yields no information about false positives. Consequently, less precise methods that tend to predict more edges will have an advantage when only raw numbers of true positives are compared, since higher false-positive rates of these tools are not considered.
To circumvent this issue and to perform a meaningful benchmark, we therefore compared methods in terms of how highly they ranked literature interactions amongst their 2000 strongest reported associations (Fig. 2 B). The underlying assumption was that methods which rank known interactions more highly will generally report more reliable relationships. To make computation feasible for all methods, we reduced the TARA Oceans data set to only OTUs that participate in at least one literature interaction. FlashWeave-S found on average 24% more literature interactions among high-ranking edges than the closest follow-up method (SpiecEasi-MB), 38% more than FlashWeave-F and on average 80% more than other methods. While the TARA Oceans data set shows considerable heterogeneity, FlashWeaveHE was not applicable due to insufficient statistical power (only 22 - 335 predicted edges total).

Pronounced runtime decreases in the Human Microbiome Project and TARA Oceans data sets

We benchmarked the computational speed of all methods on the Human Microbiome Project (HMP (The Human Microbiome Project Consortium 2012)) and TARA Oceans (Lima-Mendez et al. 2015) data sets in two settings: homogeneous and heterogeneous. For the homogeneous test, we used 2500 oral samples from the HMP data set and measured runtime on sets of 500, 750 and 1000 randomly selected OTUs (Fig. 2 A). FlashWeave outperformed other methods by factors of 8 to 158 on this benchmark (mean: 67), excluding multiple methods (SpiecEasi-GL, CoNet, mLDM) that did not finish after two days of computation (factor > 339). FlashWeave-S had on average 33% increased runtime than FlashWeave-F.

On the TARA Oceans data set (289 samples, 3762 OTUs), FlashWeave-F was on average 29 times faster than the closest non-FlashWeave method (SpiecEasi-MB), while all remaining methods did not finish computation (factor > 106; Fig. 2 A). FlashWeave-S required 53% more runtime than FlashWeave-F on this benchmark.
For the heterogeneous test, we measured the computational speed of FlashWeaveHE and all previous methods on the five body sites from the HMP data set (5514 samples, 1521 OTUs). FlashWeaveHE-F was 51 times faster than the closest non-FlashWeave method (SpiecEasi-MB) in this test and on average 371 times faster than standard FlashWeave (other methods did not finish; factor > 518; Fig. 3 C). FlashWeaveHE-S required 116% more runtime than FlashWeaveHE-F in this benchmark.
To test computational scalability in a more demanding setting, we used FlashWeaveHE-F to infer a large-scale ecological network based on 504'647 sequencing samples spanning various habitats and conditions, mapped to 75'516 OTUs at 98% 16S rRNA identity. Inference of the full interaction network completed after 1d10h46min on a High Performance Computing cluster with 200 CPU cores.

Meta variables are central hubs in the Human Microbiome Project network with high explanatory power

Meta variables (MVs), such as habitats, conditions (e.g. antibiotic usage) and technical factors (e.g. amplicon or whole-genome-shotgun sequencing) can lead to spurious associations between OTUs associated with the same MV. In addition, direct associations between MVs and OTUs can be interesting when investigating which OTUs are for instance directly associated to a particular habitat (independent of microbial interaction partners), prefer certain temperatures or are affected by specific sequencing biases.
We investigated the importance of MVs in the HMP data set by explicitly providing all five body sites and the two used primer sets (V13 vs. V35) as MVs to all FlashWeave modes. MVs formed central hubs in the resulting interaction network with on average 7.4 times larger neighborhoods than OTUs (Fig. S1 C) and 27.6 times higher betweenness centrality, a measure of node importance in the network, across all modes.
Furthermore, MVs participated in excluding up to 41.7% indirect OTU-OTU interactions (Fig. S1 B) while constituting only 0.4% of all variables. When MVs were omitted, overall numbers of OTU-OTU interactions however increased only moderately (up to 12%), suggesting that FlashWeave was generally able to use OTUs highly associated to the omitted MVs to exclude the same indirect associations. Nonetheless, when only comparing associations in direct neighborhoods of MVs, we detected 13% - 294% additional OTU-OTU associations when MVs were not provided (Fig. S1 D), indicating that MV omission may still lead to increased local biases. In addition, we found a weak association between shared primer bias and interaction probability (mean Pearson's $r < 0.003$, $P < 0.01$), suggesting only limited influence of primer preference on reported interactions. This correlation increased marginally when omitting primer information (mean $r < 0.007$, $P < 0.01$). In contrast, the univariate network showed a noticeably stronger association (mean $r < 0.057$, $P < 0.01$), suggesting less robustness to primer biases than observed for direct interaction networks.

FlashWeaveHE shows robustness to hidden meta variables and structural zeroes

While the usage of MVs can reduce the number of predicted false-positive associations, information on these variables is frequently not available because not all important latent factors are known, measured or made available in standardized annotation formats. This particularly affects inherently more heterogeneous cross-study data sets, which can feature diverse experimental, physicochemical or geographical variables, and tend to have less consistent metadata annotations.

One type of artificial associations arises from structural zeroes (Fig. 3 A), i.e. non-random absences due to unmeasured MVs. Structural zeroes can for instance occur when a data set includes multiple habitats with partially exclusive microbial content or multiple sequencing protocols biased towards disjoint OTU sets.

To compare the robustness of different methods to such absences, we computed interaction networks separately for each method and body site in the HMP data set. We then quantified the overlap of inferred interactions with a network computed on the aggregated data set of all body sites, restricted to site-specific OTUs (Fig. 3 B). We found that FlashWeaveHE showed optimal robustness to increased structural zeroes in the cross-site network, with a mean Jaccard overlap between site-specific and cross-site networks of 1.0. In contrast, homogeneous FlashWeave (0.39) and other methods (0.18 - 0.24) were markedly less robust.

Dependent sample groups constitute another type of hidden MVs, for instance re-sequencings of the same sample material with different protocols. While such groups can provide important information for network inference, for instance if certain associations can only be detected in specific experimental setups, they also break the independence assumption of common statistical association tests. We tested the impact of dependent sample groups on false positive predictions with FlashWeave through a set of simulated OTU tables with varying degrees of dependence between samples. As expected, we found that univariate networks produced by FlashWeave result in high numbers of false positive predictions when dependent samples are highly similar and constitute large fractions of a data set (Fig. S2 A). However, when computing conditional networks with FlashWeave, numbers of false positives were reduced by a median of 80% for identical samples (zero distance), with particularly strong reductions for FlashWeave-F and FlashWeaveHE-F (95%). Similarly, when increasing inter-sample distance, numbers of false positive edges in all networks dropped by medians between 89% (distance 0.25) and 99% (distance 0.75).

A large-scale interaction network from globally distributed human gut samples recovers previously described patterns and provides novel insights

We applied FlashWeaveHE to a data set of 69'818 globally distributed human gut samples ("Global Gut", GG) obtained from the NCBI Sequence Read Archive database (SRA (Leinonen et al. 2011)). The data set spanned 488 studies, the majority of which where smaller studies with less than 1000 samples (61% of all samples, 98% of all studies; Fig. S3 A). We processed samples uniformly (see Methods) and extracted sequencing protocol information and metadata

keywords from SRA annotations, resulting in a final data set of 10'624 OTUs (98% identity) and 96 MVs.
We used FlashWeaveHE to infer an interaction network (GGNcond) from GG in 3h53min using 20 CPU cores on an Intel Xeon E7-4870 machine (2.4 GHz). The method identified 30'342 significant direct interactions between OTUs and 13'151 between OTUs and MVs (30%). In contrast, when restricting FlashWeaveHE to compute a univariate network (GGNuni) we observed strongly increased edge density at 1'056'262 edges overall, 95.9% of which were excluded as indirect in GGNcond. When breaking associations in GG via shuffling (Lima-Mendez et al. 2015), FlashWeaveHE furthermore reported no false positive direct interactions. In addition, we found no evidence of dependent sample groups negatively impacting GGNcond (see Text S1).
Analyzing the American Gut Project (AGP (McDonald et al. 2018)) subset of GG (8897 samples out of 69'818) yielded a 94% decrease in predicted interactions (Fig. 4 D). For 81% of these, at least one interaction partner was absent in the AGP data set and these missing partners tended to be rare in GGNcond, with 87% decreased mean prevalence in GG compared to OTUs found in both data sets.
We found the OTU-OTU sub-network of GGNcond to be strongly structured (modularity 0.25), indicating the presence of distinct communities. The 20 largest clusters had on average 45 members (up to 89) and featured almost exclusively positive interactions between members (mean 99.6%), but only 37.1% - 79.8% (mean 63.3%) positive edges to non-member OTUs. Similarly, we found the majority of positive interactions per phylum to be within-phylum interactions (50% in *Actinobacteria* and up to 87% in *Firmicutes*, mean 68%), while negative interactions frequently featured partners from other phyla (35% in *Firmicutes* up to 95% in *Actinobacteria*, mean 73%). For *Actinobacteria*, which had the highest fraction of negative edges to other phyla, the majority targeted *Firmicutes* (48%) and *Bacteroides* (35%).
Many negative interactions in GGNcond were mediated by a few dominant OTUs (Fig. 4 A, C), which constituted negative hubs not explainable by our set of MVs (see Methods). These include several species implied in inflammation and disease (*Dorea formicigenerans (Guinane & Cotter 2013)*, *Bilophila wadsworthia* (Feng et al. 2017), *Odoribacter splanchnicus (Werner et al. 1975)*, *Bacteroides vulgatus (Ó Cuív et al. 2017)*). Additionally, we found negative associations between multiple *Blautia* OTUs and a *Clostridium difficile* OTU, consistent with previous reports (Daquigan et al. 2017; Stein et al. 2013).
Phylogenetic assortativity (PA), i.e. the increased probability of interaction between evolutionarily less diverged partners, is a frequently observed ecological pattern of potential biological interest (Chaffron et al. 2010; Faust et al. 2012; Kurtz et al. 2015). We found pronounced PA in GGNcond for positive edges, while negative edges were closer to the empirical null distribution (Fig. 4 B, lower row). Though differences were significant in both cases (two-sample Kolmogorov-Smirnov test, $P < 0.01$), effect size was increased by 10x for positive edges. In contrast, positive edges in GGNuni showed a noticeably smaller effect size increase over negative edges (3.7x increase, Fig. 4 B, upper row).
Among OTUs with the highest numbers of positive neighbors (Fig. S3 D), constituting potential candidates for keystone species (Berry & Widder 2014), we observed several OTUs from *Bacteroides* (genus) and numerous *Clostridiales* (order) OTUs, both taxa known to harbor important mutualist species in the human gut (Fischbach & Sonnenburg 2011; Lopetuso et al.

2013). Intriguingly, 75% of the top 20 positive hubs were taxonomically uncharacterized at the genus level.
Consistent with known dependencies between H_2 producing and consuming microbes which have been described in the human gut (Carbonero et al. 2012), we found significantly more positive interactions between H_2 producers and consumers in GGNcond than in random networks, accounting for PA as a possible confounder (3.6x increase, empirical $P < 0.01$). This effect was noticeably weaker for GGNuni (1.8x increase, empirical $P < 0.01$).

Discussion

FlashWeave is the first ecological network inference approach that combines i) the estimation of direct interactions, ii) the ability to scale to large-scale data sets with tens of thousands of OTUs and hundreds of thousands of samples and iii) the incorporation of meta variable (MV) information.
We showed that FlashWeave typically outperforms state-of-the-art methods by several orders of magnitude in terms of runtime, compares favourably at recovering gold-standard literature interactions in the TARA Oceans data set and generally surpasses other methods in terms of edge accuracy on a wide selection of synthetic benchmarks. Notably, network recovery typically continued improving for FlashWeave with additional samples, indicating effective utilization of additional power provided by larger data sets, which are becoming increasingly available.
Sequenced samples do not only increase in number, but also in heterogeneity, as more habitats are being sampled under a plethora of different conditions and experimental protocols. These factors may confound association signals, resulting in biased interaction networks. FlashWeave tackles this challenge in two ways: firstly, it features a specialized heterogeneous data mode (FlashWeaveHE) which, as we show, achieves strongly improved consistency, edge accuracy and runtime compared to other methods in the presence of structural zeros, which can be problematic in data sets with high sample heterogeneity (Röttjers & Faust 2018). Secondly, FlashWeave can use MV information to exclude spurious OTU-OTU associations driven by these MVs. Exemplified by primers and body sites in the HMP data set, we observed that omission of MVs resulted in noticeable increases of edge density between OTUs directly associated with these variables, analogous to spurious edges induced between neighbors of keystone taxa (Berry & Widder 2014). Interestingly, we found FlashWeave predictions to still be remarkably robust to the omission of MV information when considering the HMP network as a whole, indicating that relatively accurate interaction networks may be inferrable even in the absence of MV information. This finding is further supported by FlashWeave's robustness to dependent sample groups which we observed both in simulations and in the Global Gut data set. However, a more comprehensive analysis would be required in the future to investigate the prospects and limits of this effect in more detail.
Besides supporting exclusion of spurious OTU-OTU associations, direct relationships between OTUs and MVs reported by FlashWeave furthermore yield insights into non-microbial factors influencing microbial ecosystems. Exemplified by the HMP analysis, we found MVs to be central nodes in the association network with many directly associated OTUs, in line with the expected high dependence of many microbes on specific habitats (The Human Microbiome Project

Consortium 2012) and the known existence of primer biases (Tremblay et al. 2015). Consistent with our results, closely related approaches have previously been used to identify parsimonious sets of highly predictive microbial biomarkers for human body sites and a skin disease (Tackmann et al. 2018; Statnikov et al. 2013).

We demonstrated that FlashWeave scales to modern data sets by computing networks for two aggregated cross-study data sets with 69 818 human gut samples ("Global Gut", GG) and 504 647 multi-habitat samples, which finished computation in less than 4h and 1.5d, respectively. We found that using GG for network construction strongly improved the number of predicted interactions compared to a network based on the GG subset corresponding to the American Gut Project (AGP). This result indicates that predictions based on single-study data sets may miss large fractions of associations. The majority of missing edges are likely due to lack of statistical power, since networks for random size-matched subsets of GG also displayed markedly reduced edge numbers. However, these networks were still noticeably larger than the AGP network, indicating that increased data heterogeneity also benefits edge detection. The vast majority of missed interactions involved low-prevalence OTUs, highlighting that more comprehensive data sets such as GG may allow first glimpses at ecological interactions of the hitherto underexplored rare microbial biosphere (Yang et al. 2017; Jousset et al. 2017). This is a crucial advancement, as most analyses to date are restricted to highly prevalent OTUs, due to lack of statistical power or computational limitations.

Encouragingly, the FlashWeave network inferred for the GG data set (GGNcond) furthermore revealed consistency with several expected biological patterns, such as a known dependency between H_2 producers and consumers (Carbonero et al. 2012), mutualist hubs mapping to taxa associated with cross-feeding (Fischbach & Sonnenburg 2011; Lopetuso et al. 2013), a previously reported negative interaction partner of *Clostridium difficile (Daquigan et al. 2017; Stein et al. 2013)*, and phylogenetic assortativity of interacting microbes (Chaffron et al. 2010; Faust et al. 2012; Kurtz et al. 2015).

GGNcond featured several hub OTUs with mainly negative interactions, a number of which map to disease or dysbiosis-associated species. While we cannot fully rule out potential indirect influences of host-related factors - albeit we made every effort to avoid this possibility - these hubs nonetheless provide a valuable list of candidates for further experimental validation. In particular, elucidating potential ecological mechanisms (e.g. wide-spread competitive repression or antagonism) driving the negative impact that these OTUs may have on the gut ecosystem would be intriguing.

Furthermore, we found that the vast majority of the most distinct positive hub OTUs in GGNcond were not confidently classifiable at the genus level, indicating a crucial lack of information on potentially important mutualists in the human gut. A remarkable number of these were assigned to the families *Lachnospiraceae* and *Ruminococcaceae* (order *Clostridiales*), providing evidence that the positive role of unclassified OTUs from these families on ecosystem maintenance may be more pronounced than currently appreciated (Lopetuso et al. 2013).

The importance of excluding indirect associations has been pointed out previously (Kurtz et al. 2015; Yang et al. 2017; Röttjers & Faust 2018) and we found clear advantages of doing so in GGNcond: both phylogenetic assortativity (PA) and H_2 producer/consumer signals were noticeably more pronounced in GGNcond compared to a univariate network with both direct and indirect associations (GGNuni). Strikingly, GGNcond featured 96% fewer associations than

GGNuni, mirroring results from our synthetic benchmarks and suggesting substantial increases in false positives in methods that don't account for indirect associations. However, we note that more data is typically needed to fully support the removal of indirect associations. Furthermore, niche-driven indirect interactions may still increase resolution in community structure indices that explicitly incorporate co-occurrence information (Schmidt et al. 2017). PA signals can be confounded by shared niche preference, because phylogenetically more closely related organisms tend to have similar niches, and this may explain part of the PA signal we observed in GGNuni. However, GGNcond is specifically depleted for such indirect associations, yet shows a stronger PA signal than GGNuni. This suggests that the observed PA is driven by niche-independent factors. A possible explanation would be that kin selection (Strassmann et al. 2011), as previously observed for instance in biofilms (Xavier & Foster 2007) or iron acquisition (Griffin et al. 2004), is more pronounced in the human gut than currently appreciated. However, we could not entirely rule out shared-niche contributions in GGNcond and therefore deem future confirmatory investigation of this finding necessary.

Current limitations of FlashWeave include the handling structural zeros in the FlashWeaveHE modes, which currently conservatively discard any absences. While we found the resulting power reduction to be noticeable for data sets with fewer samples (TARA Oceans), this effect was mitigated for larger sample sizes in our synthetic benchmarks. Since globally distributed cross-study data sets include even more samples, and additional samples are continuously being sequenced, power issues should therefore not be a strong limitation in typical use cases. Since insufficient power may nonetheless affect recovery of interactions between rare taxa, more refined models assigning confidences to absences would be interesting additions in future versions of FlashWeaveHE. Another interesting aspect that remains to be explored in more detail is the impact of MVs on a broader range of studies and variables, such as marine physiological factors or human disease conditions.

The LGL framework, which FlashWeave builds upon, is highly flexible and permits several straightforward extensions, such as more powerful tests (Xu et al. 2015; Lovell et al. 2015) and more importantly the prediction of edge directionality (Aliferis et al. 2010a). The latter is an exciting prospect that would enable a more causal interpretation of predicted ecological interactions, paving the path towards efficiently learning fully predictive models. In the future, such data-driven models may allow us to forecast the ecological impact of perturbations and catalyze emerging ecological engineering applications.

Author contributions

Conceptualization, J.T., J.F.M.R. and C.v.M.; Methodology, J.T. and J.F.M.R.; Software, J.T.; Investigation, J.T., J.F.M.R. and C.v.M.; Writing – Original Draft, J.T.; Writing – Review & Editing, J.T., J.F.M.R. and C.v.M.; Visualization, J.T.; Funding Acquisition, C.v.M.; Supervision, J.F.M.R. and C.v.M.

Acknowledgments

We thank Jeroen Raes, Shinichi Sunagawa and Reinhard Furrer for their valuable methodological feedback and Maria Dmitrieva for helpful discussions during the preparation of this manuscript. This work was supported by the Swiss National Science Foundation (grant nr. 31003A-160095).

Declaration of interests

The authors declare no competing interests.

Methods

The Algorithm

FlashWeave is implemented in the Julia Programming Language (Bezanson et al. 2017) and based on the local-to-global learning framework (LGL (Aliferis et al. 2010a)). Causal inference algorithms of this class start by performing a locally optimal Markov blanket search in order to infer all directly associated neighbors of a target variable T (OTU or MV in the case of FlashWeave), representing the set of estimated direct causes and effects of T. Then, individual neighborhoods are connected through a combinator rule (by default the OR rule in FlashWeave) to form a global association graph. In the final step, currently not implemented in FlashWeave, this undirected skeleton of conditional dependence relationships can be used as a scaffold to efficiently infer edge directionality and provide further insights into the system of study. In line with results in (Aliferis et al. 2010b), FlashWeave employs a False Discovery Rate (FDR) adjustment step and omits the costly steps of spouse identification and symmetry correction.

LGL can be instantiated with a wide range of algorithms and conditional independence tests. FlashWeave currently defaults to the efficient semi-interleaved HITON-PC algorithm (Aliferis et al. 2010a) and provides the choice of either discretized mutual information tests (more coarse grained and usually quicker; "fast" mode) or partial correlation tests (more sensitive and usually slower; "sensitive" mode) (Fig. 1 C, Fig. S1 A).

FlashWeaveHE further specializes these tests to exclude zero elements from association computations (Fig. 1 C, Fig. S1 A). It makes the assumption that zeroes in large, heterogeneous data sets are mostly structural (for instance due to primer or sequencing depth biases, as well as habitat or condition-specific effects) and thereby only considers samples in which both OTUs have a non-zero abundance as reliable for association prediction. Notably, this restriction is only applied to the prospective interaction partners being tested: OTUs found in the conditioning set retain their absences. This procedure is chosen i) to not discard too much information concerning the tested partners, which would otherwise result in drastic loss of power as conditioning sets grow, and ii) because structural absences of OTUs within conditioning sets, despite marginally decreasing power, otherwise have minimal impact on exclusion decisions. While the

FlashWeaveHE approach potentially discards some valid absence information and can thereby be less sensitive than the vanilla mode of FlashWeave, we found this loss in sensitivity to be small in heterogeneous data sets with larger sample sizes. Indeed, FlashWeaveHE resulted in strongly increased precision (Fig. 3 B, D) and much improved runtimes (Fig. 3 C) on such data.
Normalization in FlashWeave accounts for compositionality effects and differs depending on the test type. Details on normalization schemes and a discussion of novel heuristics can be found in Text S1.

Accuracy and robustness benchmarks

For the NorTA (American Gut) benchmark, synthetic data was generated as described in (Kurtz et al. 2015) using the "amgut.filt" data set, "cluster" and "scale free" topologies and default settings for all other parameters. In order to increase the compositionality signal, we downsampled each sample to depths randomly picked from "amgut.filt".
For the Ecological Models benchmark, data sets were generated as described in (Weiss et al. 2016), restricted to linear ecological relationships. Three independent data sets (500, 1000 and 2000 samples) were generated per table. To create data sets with multiple disjoint habitats (Fig. 3 D), the three differently sized data sets per table were aggregated with OTUs and interactions assumed as distinct, i.e. with each OTU and interaction only present in one habitat.
For the structural zero robustness benchmark, we reduced all body sites in the HMP data set via random subsampling to a fixed number of 312 samples per site. For each body site, we then picked all OTUs found in at least 10 samples of that site (175 - 619 OTUs) and removed their non-zero counts from all samples belonging to other body sites. The resulting body site-specific data sets were then aggregated into a single table. Inference tools were applied to i) each individual body site table separately, ii) the aggregated data set of all body sites. Finally, the edge overlaps between all sub-networks and the aggregated network were compared using the Jaccard similarity index.
Dependent sample groups were simulated in the following fashion: First, a dependence-free data set was simulated using a zero-inflated multivariate log-normal distribution, constructed with the Distributions.jl package (JuliaStats 2018a). OTUs were simulated as ecologically independent (covariance matrix = identity matrix), a vector of log-means for the log-normal component was sampled uniformly from range 2 to 10 and parameters for the zero-inflated multinomial component were sampled from a Beta distribution with $\alpha = 1$ and $\beta = 3$. This model M was used to simulate abundances for 200 OTUs in 10'000 samples, resulting in OTU table A_{ind}. In addition to this dependence-free data set, sets S_n^f of dependent sample groups $g_{i,n}^f$ were generated, where $n \in \{5, 50, 100\}$ was the number of dependent sample groups per set, $i \in \{1 \,..\, n\}$ was the group index and $f \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ was the fraction of samples in A_{ind} to be replaced with S_n^f (i.e. the dependence fraction). Additionally, we simulated different within-group distances $d \in \{0.0, 0.25, 0.5, 0.75\}$ for each $g_{i,n}^f$ through constrained iterative sampling, where new samples were generated from Muntil the desired f was reached and new samples were only accepted when the mean Bray-Curtis distance to all previously accepted samples was within 0.01 from d. In the special case of $d = 0.0$, random samples from A_{ind} were picked and repeated within $g_{i,n}^f$

until f was reached. Empirical distances in the simulated sample groups closely matched target distances (Fig. S2 B).

Table S1

Parameters and software versions used for each inference method. All other parameters were kept at default values.

Method	Version	Parameters
FlashWeave	0.9.0	max-k: 3, conv: 0.01, feed-forward & fast-elimination enabled
SpiecEasi	0.1.3	lambda.min.ratio: 1e-2, nlambda: 15, nrep: 20
eLSA	81a2ee0	p: theo, n: percentileZ, f: zero, q-value < 0.05
SparCC	See SpiecEasi	q-value < 0.05, FDR correction via p.adjust (option "BH") in R (R Core Team 2017)
CoNet	1.1.1	methods: correl_spearman/dist_kullbackleibler/dist_bray, multitestcorr: benjaminihochberg, q-value < 0.05
mLDM	1.0	Z_mean: 1

Literature interaction predictions

To reduce computation time, we filtered the TARA Oceans data set down to only OTUs participating in at least one genus-level literature interaction reported in (Lima-Mendez et al. 2015). After removing samples with no reads, the final data set consisted of 234 samples and 702 OTUs. Edges predicted by each tool were sorted according to their reported weights (merged q-value in the case of CoNet, which uses multiple weight measures; Pearson's *r* for SparCC) and this ranking was plotted as cumulative curves (Fig. 2 B).

Computational speed benchmarks

The HMP data set consisted of 5514 samples from the body sites oral, gastrointestinal tract, urogenital tract, skin and airways. Samples were mapped to OTUs at 96% 16S rRNA identity using MAPseq (version v1.0 (Matias Rodrigues et al. 2017), confidence > 0.5) and the full-length 16S reference provided with MAPseq. The data set was further filtered for OTUs present in more than 20 samples.

For the TARA Ocean, we aggregated the preprocessed OTU counts tables provided by (Lima-Mendez et al. 2015) into a single data set. After filtering for OTUs present in more than 50 samples and samples with at least one read, the data set contained 289 samples and 3762 OTUs.

Parameters for each network inference tool were as reported in Table S1. Since not all tools readily supported parallelism, benchmarks were conducted on a single core on an AMD Opteron 2347 HE machine (1 GHz).

Meta variable analysis in the HMP

An indirect association was counted as explained by a MV if at least one MV was present in the set of conditional variables leading to the association's exclusion (Fig. S1 B). The correlation between shared primer influence and interaction probability was estimated by computing, for each pair of OTUs (O_i, O_j), the absolute difference of association strengths in the HMP network between O_i and the primer MV and O_j and the primer MV, leading to small values for OTU pairs with similar primer influence and larger values for differences in influence. These values were then correlated with the interaction strengths between each O_i and O_j using Pearson's *r*.

Global Gut network analysis

Data set creation and network computation

Studies from the NCBI Sequence Read Archive database (SRA (Leinonen et al. 2011)) were filtered for human samples through the automated parsing of metadata annotation fields, matching at least one of the following rules: 1) "Human" or "Homo sapiens" is found in the host name field, 2) "9606" is found in either the host taxon ID or sample taxon ID field, or 3) "human (gut|gastrointestinal) metagenome" is found in the organism field, where "(gut|gastrointestinal)" is a regular expression match for either "gut" or "gastrointestinal". For matching samples, a list of keywords was parsed from all main annotation fields and further curated to remove uninformative terms, resulting in a set of keywords assigned to each sample. Samples were then further filtered for gut association by only retaining samples matching at least one of the following keywords: "intestinal", "intestine", "alimentary", "bowel", "cecum", "crohn", "gut", "colon", "commensal-gut", "diarrhoea", "digestive-tract", "digestive tract", "duodenum", "enteric", "enteritis", "enterocolitis", "enteropathogenic", "enterohemorrhagic", "equol", "feces", "gastroenteritis", "gastrointestinal", "ileum", "ileostomy", "jejunum", "meconium", "mesentery", "mid-gut", "probiotic", "rectum", "stec", "vibriosis". Keywords of these samples were additionally checked for terms not related to gut, followed by manual review of such samples via the SRA web service and removal in case of likely non-gut origin.

The final set of samples was downloaded and mapped to OTUs at 98% 16S rRNA identity using MAPseq (version v1.0 (Matias Rodrigues et al. 2017), confidence > 0.5) and the full-length 16S rRNA reference database provided with MAPseq (hierarchically clustered with HPC-CLUST (Matias Rodrigues & von Mering 2014) using average linkage). We removed samples with less than 100 mapped reads and OTUs found in less than 200 samples (see Table S2 for SRA accessions of the final sample set). Taxonomy was assigned to OTUs based on a 90% consensus over the full taxonomic lineages of all OTU member sequences. A trusted set of taxonomic

classifications was generated using the annotated taxonomy provided by NCBI (updated on February 2018) including only sequences belonging to RefSeq (O'Leary et al. 2016) genomes and sequences from culture collection strains. The remaining sequences were taxonomically classified using a version of MAPseq modified to compute global alignments and the trusted set of sequences with their associated taxonomies (confidence cutoff >= 0.5). Applied identity cutoff and scaling parameters (delimited by colons) were 0.00:0.08 (Kingdom), 0.75:0.035 (Phylum), 0.785:0.035 (Class), 0.82:0.045 (Order), 0.865:0.06 (Family), 0.92:0.06 (Genus), 0.95:0.05 (Species), with identity cutoffs as suggested in (Yarza et al. 2014).
In addition, we retrieved sequencing method information from the SRA ("WGS", "AMPLICON" "RNA-SEQ" or "OTHER") and filtered the previously extracted metadata keywords for a set of 128 potentially interesting terms such as "fibre", "antibiotics" and "cancer". This metadata information was used to create a MV information table which was further hierarchically clustered into 96 MV groups (average linkage, unweighted Jaccard similarity > 0.9). See Table S3 for representatives picked for each group.
The OTU table and the MV group table were finally used as input to FlashWeaveHE-F with parameters reported in Table S1 to compute the GGNcond and AGP networks and with max-k = 0 to compute GGNuni.

FDR estimation and modularity

To estimate the false positive rate, we generated a null model by breaking associations between taxa through sequencing depth-conserving shuffling of the GG data set (Lima-Mendez et al. 2015). Modularity (Newman 2006) was computed based on cluster assignments from Markov Chain Clustering (MCL (Van Dongen 2008) version 14-137, inflation parameter 1.5).

Influence of dependent sample groups on GGNcond

For computational efficiency, the GG data set was clustered using an iterative greedy approach, in which samples were initially sorted by number of mapped reads (descending order). Iterating through that order, the clustering algorithm checked in each step if any subsequent samples were within the desired distance threshold d and added these samples to the current cluster, removing them from future consideration. The procedure was repeated for $d \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ and for each clustering, a random downsampling of the GG data set with number of samples equal to the corresponding number of clusters was performed to create separate background data sets. Univariate and conditional networks for each clustering and background data set were computed using the same parameters as used for GGNcond and GGNuni, respectively.

Impact of meta variables on negative hubs

In order to estimate whether negative associations of the top 20 negative hub OTUs could be explained by MVs, we collected all negative associations of these OTUs and computed for each MV, how often samples assigned with this MV contributed to a positive or negative association signal within the negative edges. We then compared MV frequencies of negative contributions to those of positive contributions and found no significant difference (paired T-test, P > 0.99),

indicating that positive and negative association signals were overall driven by samples with highly similar MV distributions.

Phylogenetic assortativity

For phylogenetic tree construction, the alignment of representatives for all 98% 16S rRNA identity OTUs in the MAPseq reference database (92'659 full-length 16S rRNA sequences) was computed with INFERNAL (version 1.1.2 (Nawrocki & Eddy 2013)) using the microbial secondary structure model SSU-ALIGN (Nawrocki & Eddy 2013). The phylogenetic tree was then reconstructed using fasttree (version 2.1.3 (Price et al. 2010)) with the GTR substitution model and otherwise default options. For the phylogenetic assortativity analysis, the GGNcond network was reduced into two separate networks restricted to edges and vertices participating in only positive and negative associations, respectively. To generate a random background, vertices in each network were randomly connected to create a network with vertex and edge numbers matching the original network. Phylogenetic distance between interaction partners was calculated as total branch length between the leaves corresponding to these OTUs. The same procedure was repeated for GGNuni to estimate phylogenetic assortativity of univariate edges.

Associations between H_2 producers and consumers

OTU that mapped to H_2 producing and consuming taxa (taken from (Carbonero et al. 2012)) were identified in GGNcond. The number of positive interactions between these groups was compared to interactions between the same groups in 100 randomly generated networks. To assure comparability, random networks were generated such that the expected positive degree for each OTU was conserved and the interaction probabilities respected the phylogenetic assortativity signal detected in GGNcond. The latter was done to assure that non-random interaction patterns were not explainable by phylogenetic assortativity alone. This step was implemented by using the package KernelDensity.jl (JuliaStats 2018b) to fit a Kernel Density Estimate (Gaussian Kernel with $\mu = 0.0$ and $\sigma = 0.25$) per OTU O_i to the phylogenetic distances D_{ij} between O_i of all of its positive interaction partners O_j^i, which yielded distribution P_i. When sampling neighbors of O_i, the probability π_{ij} of OTUs O_i and O_j interacting was then computed as the reciprocal product $P_i(D_{ij}) \cdot P_j(D_{ij})$, followed by re-normalization of $\pi_i = \{\pi_{ij},\ j \neq i\}$ to a proper probability density. Degree conservation was achieved by only considering OTUs O_j for interaction if their current degree was still smaller than in GGNcond.

Normalization comparison

The subset of Gastrointestinal tract samples from the HMP data set was filtered along sequencing depth and OTU prevalence gradients, followed by applying *clr* (pseudo-count 1) and *clr-adapt* normalization schemes (Text S1). Associations were inferred using FlashWeave-S with max-k 0 (univariate) and max-k 3 (conditional) and all other options as in Table S1. For the oral comparison, the oral subset of the HMP data set (1000 OTUs) was used and networks were computed with 20 CPU cores.

Data and Software Availability

FlashWeave is open source software implemented in Julia (Bezanson et al. 2017) and freely available from https://github.com/meringlab/FlashWeave.jl under the GNU General Public License v3.0.

References

Aitchison, J., 1981. A new approach to null correlations of proportions. *Journal of the International Association for Mathematical Geology*, 13(2), pp.175–189.

Aliferis, C.F. et al., 2010a. Local causal and markov blanket induction for causal discovery and feature selection for classification part i: Algorithms and empirical evaluation. *Journal of machine learning research: JMLR*, 11(1), pp.171–234.

Aliferis, C.F. et al., 2010b. Local causal and markov blanket induction for causal discovery and feature selection for classification part ii: Analysis and extensions. *Journal of machine learning research: JMLR*, 11(1), pp.235–284.

Berry, D. & Widder, S., 2014. Deciphering microbial interactions and detecting keystone species with co-occurrence networks. *Frontiers in microbiology*, 5, p.219.

Bezanson, J. et al., 2017. Julia: A Fresh Approach to Numerical Computing. *SIAM Review*, 59(1), pp.65–98.

Bonder, M.J. et al., 2016. The effect of host genetics on the gut microbiome. *Nature genetics*, 48(11), pp.1407–1412.

Carabotti, M. et al., 2015. The gut-brain axis: interactions between enteric microbiota, central and enteric nervous systems. *Annales de gastroenterologie et d'hepatologie*, 28(2), pp.203–209.

Carbonero, F., Benefiel, A.C. & Gaskins, H.R., 2012. Contributions of the microbial hydrogen economy to colonic homeostasis. *Nature reviews. Gastroenterology & hepatology*, 9(9), pp.504–518.

Chaffron, S. et al., 2010. A global network of coexisting microbes from environmental and whole-genome sequence data. *Genome research*, 20(7), pp.947–959.

Daquigan, N. et al., 2017. High-resolution profiling of the gut microbiome reveals the extent of Clostridium difficile burden. *npj Biofilms and Microbiomes*, 3(1). Available at: http://dx.doi.org/10.1038/s41522-017-0043-0.

Duda, S. et al., 2005. Extracting drug-drug interaction articles from MEDLINE to improve the content of drug databases. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, pp.216–220.

Dyhrman, S., Ammerman, J. & Van Mooy, B., 2007. Microbes and the Marine Phosphorus Cycle. *Oceanography* , 20(2), pp.110–116.

Falkowski, P.G., Fenchel, T. & Delong, E.F., 2008. The Microbial Engines That Drive Earth's Biogeochemical Cycles. *Science*, 320(5879), pp.1034–1039.

Faust, K. et al., 2012. Microbial Co-occurrence Relationships in the Human Microbiome. *PLoS computational biology*, 8(7), p.e1002606.

Faust, K. & Raes, J., 2016. CoNet app: inference of biological association networks using Cytoscape. *F1000Research*, 5, p.1519.

Faust, K. & Raes, J., 2012. Microbial interactions: from networks to models. *Nature reviews. Microbiology*, 10(8), pp.538–550.

Feng, Z. et al., 2017. A human stool-derived Bilophila wadsworthia strain caused systemic inflammation in specific-pathogen-free mice. *Gut pathogens*, 9(1). Available at: http://dx.doi.org/10.1186/s13099-017-0208-7.

Fischbach, M.A. & Sonnenburg, J.L., 2011. Eating for two: how metabolism establishes interspecies interactions in the gut. *Cell host & microbe*, 10(4), pp.336–347.

Friedman, J. & Alm, E.J., 2012. Inferring correlation networks from genomic survey data. *PLoS computational biology*, 8(9), p.e1002687.

Goers, L., Freemont, P. & Polizzi, K.M., 2014. Co-culture systems and technologies: taking synthetic biology to the next level. *Journal of the Royal Society, Interface / the Royal Society*, 11(96). Available at: http://dx.doi.org/10.1098/rsif.2014.0065.

Griffin, A.S., West, S.A. & Buckling, A., 2004. Cooperation and competition in pathogenic bacteria. *Nature*, 430(7003), pp.1024–1027.

Guinane, C.M. & Cotter, P.D., 2013. Role of the gut microbiota in health and chronic gastrointestinal disease: understanding a hidden metabolic organ. *Therapeutic advances in gastroenterology*, 6(4), pp.295–308.

Jousset, A. et al., 2017. Where less may be more: how the rare biosphere pulls ecosystems strings. *The ISME journal*, 11(4), pp.853–862.

JuliaStats, 2018a. KernelDensity.jl, Kernel density estimators for Julia. Available at: https://github.com/JuliaStats/KernelDensity.jl.

JuliaStats, 2018b. KernelDensity.jl, Kernel density estimators for Julia. Available at: https://github.com/JuliaStats/KernelDensity.jl.

Kawaguchi, M. & Minamisawa, K., 2010. Plant-microbe communications for symbiosis. *Plant & cell physiology*, 51(9), pp.1377–1380.

Krause, E. et al., 2012. Small changes in pH have direct effects on marine bacterial community composition: a microcosm approach. *PloS one*, 7(10), p.e47035.

Kurtz, Z.D. et al., 2015. Sparse and compositionally robust inference of microbial ecological networks. *PLoS computational biology*, 11(5), p.e1004226.

Leinonen, R. et al., 2011. The sequence read archive. *Nucleic acids research*, 39(Database issue), pp.D19–21.

Levy, R. et al., 2015. NetCooperate: a network-based tool for inferring host-microbe and microbe-microbe cooperation. *BMC bioinformatics*, 16, p.164.

Lima-Mendez, G. et al., 2015. Ocean plankton. Determinants of community structure in the global plankton interactome. *Science*, 348(6237), p.1262073.

Lopetuso, L.R. et al., 2013. Commensal Clostridia: leading players in the maintenance of gut homeostasis. *Gut pathogens*, 5(1), p.23.

Lovell, D. et al., 2015. Proportionality: a valid alternative to correlation for relative data. *PLoS computational biology*, 11(3), p.e1004075.

Matias Rodrigues, J.F. et al., 2017. MAPseq: highly efficient k-mer search with confidence estimates, for rRNA sequence analysis. *Bioinformatics* . Available at: http://dx.doi.org/10.1093/bioinformatics/btx517.

Matias Rodrigues, J.F. & von Mering, C., 2014. HPC-CLUST: distributed hierarchical clustering for large sets of nucleotide sequences. *Bioinformatics* , 30(2), pp.287–288.

McDonald, D. et al., 2018. American Gut: an Open Platform for Citizen Science Microbiome Research. *mSystems*, 3(3). Available at: http://dx.doi.org/10.1128/mSystems.00031-18.

McFall-Ngai, M., 2014. Divining the essence of symbiosis: insights from the squid-vibrio model. *PLoS biology*, 12(2), p.e1001783.

Mitchell, A.L. et al., 2018. EBI Metagenomics in 2017: enriching the analysis of microbial communities, from sequence reads to assemblies. *Nucleic acids research*, 46(D1), pp.D726–D735.

Narendra, V. et al., 2011. A comprehensive assessment of methods for de-novo reverse-engineering of genome-scale regulatory networks. *Genomics*, 97(1), pp.7–18.

Nawrocki, E.P. & Eddy, S.R., 2013. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* , 29(22), pp.2933–2935.

Newman, M.E.J., 2006. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23), pp.8577–8582.

Ó Cuív, P. et al., 2017. The gut bacterium and pathobiont Bacteroides vulgatus activates NF-κB in a human gut epithelial cell line in a strain and growth phase dependent manner. *Anaerobe*, 47, pp.209–217.

Oh, D.-C. et al., 2009. Dentigerumycin: a bacterial mediator of an ant-fungus symbiosis. *Nature chemical biology*, 5(6), pp.391–393.

O'Leary, N.A. et al., 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic acids research*, 44(D1), pp.D733–45.

Pearson, K., 1896. Mathematical Contributions to the Theory of Evolution.--On a Form of

Spurious Correlation Which May Arise When Indices Are Used in the Measurement of Organs. *Proceedings of the Royal Society of London*, 60(1), pp.489–498.

Price, M.N., Dehal, P.S. & Arkin, A.P., 2010. FastTree 2--approximately maximum-likelihood trees for large alignments. *PloS one*, 5(3), p.e9490.

R Core Team, 2017. R: A Language and Environment for Statistical Computing. Available at: https://www.R-project.org/.

Röttjers, L. & Faust, K., 2018. From hairballs to hypotheses-biological insights from microbial networks. *FEMS microbiology reviews*, 42(6), pp.761–780.

Sboner, A. & Aliferis, C.C.F., 2005. Modeling clinical judgment and implicit guideline compliance in the diagnosis of melanomas using machine learning. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, p.664.

Schmidt, T.S.B., Matias Rodrigues, J.F. & von Mering, C., 2017. A family of interaction-adjusted indices of community similarity. *The ISME journal*, 11(3), pp.791–807.

Solden, L., Lloyd, K. & Wrighton, K., 2016. The bright side of microbial dark matter: lessons learned from the uncultivated majority. *Current opinion in microbiology*, 31, pp.217–226.

Statnikov, A. et al., 2013. Microbiomic signatures of psoriasis: feasibility and methodology comparison. *Scientific reports*, 3, p.2620.

Stein, R.R. et al., 2013. Ecological modeling from time-series inference: insight into dynamics and stability of intestinal microbiota. *PLoS computational biology*, 9(12), p.e1003388.

Strassmann, J.E., Gilbert, O.M. & Queller, D.C., 2011. Kin discrimination and cooperation in microbes. *Annual review of microbiology*, 65, pp.349–367.

Tackmann, J. et al., 2018. Ecologically informed microbial biomarkers and accurate classification of mixed and unmixed samples in an extensive cross-study of human body sites. *Microbiome*, 6(1), p.192.

Thaiss, C.A. et al., 2016. The microbiome and innate immunity. *Nature*, 535(7610), pp.65–74.

The Human Microbiome Project Consortium, 2012. Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402), pp.207–214.

Tremblay, J. et al., 2015. Primer and platform effects on 16S rRNA tag sequencing. *Frontiers in microbiology*, 6, p.771.

Vandeputte, D. et al., 2017. Quantitative microbiome profiling links gut community variation to microbial load. *Nature*, 551(7681), pp.507–511.

Van Dongen, S., 2008. Graph Clustering Via a Discrete Uncoupling Process. *SIAM Journal on Matrix Analysis and Applications*, 30(1), pp.121–141.

Weiss, S. et al., 2016. Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *The ISME journal*, 10(7), pp.1669–1681.

Werner, H., Rintelen, G. & Kunstek-Santos, H., 1975. [A new butyric acid-producing bacteroides

species: B. splanchnicus n. sp. (author's transl)]. *Zentralblatt fur Bakteriologie, Parasitenkunde, Infektionskrankheiten und Hygiene. Erste Abteilung Originale. Reihe A: Medizinische Mikrobiologie und Parasitologie*, 231(1-3), pp.133–144.

Xavier, J.B., 2011. Social interaction in synthetic and natural microbial communities. *Molecular systems biology*, 7, p.483.

Xavier, J.B. & Foster, K.R., 2007. Cooperation and conflict in microbial biofilms. *Proceedings of the National Academy of Sciences of the United States of America*, 104(3), pp.876–881.

Xia, L.C. et al., 2011. Extended local similarity analysis (eLSA) of microbial community and other time series data with replicates. *BMC systems biology*, 5 Suppl 2, p.S15.

Xu, L. et al., 2015. Assessment and Selection of Competing Models for Zero-Inflated Microbiome Data. *PloS one*, 10(7), p.e0129606.

Yang, Y., Chen, N. & Chen, T., 2017. Inference of Environmental Factor-Microbe and Microbe-Microbe Associations from Metagenomic Data Using a Hierarchical Bayesian Statistical Model. *Cell systems*, 4(1), pp.129–137.e5.

Yarza, P. et al., 2014. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nature reviews. Microbiology*, 12(9), pp.635–645.

Zelezniak, A. et al., 2015. Metabolic dependencies drive species co-occurrence in diverse microbial communities. *Proceedings of the National Academy of Sciences of the United States of America*, 112(20), pp.6449–6454.

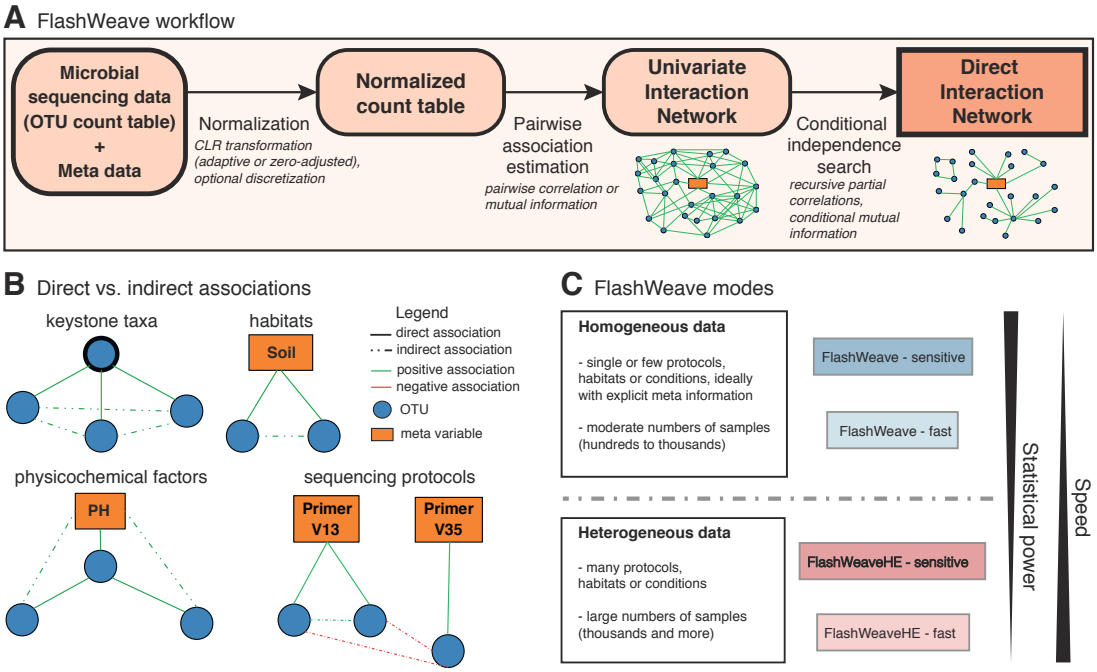


Figure 1: Overview of FlashWeave. A Main steps in the network inference pipeline. **B** Examples of how indirect associations may create false positive results in various ecological scenarios or for different experimental protocols. **C** Use cases of the different modes available for FlashWeave.

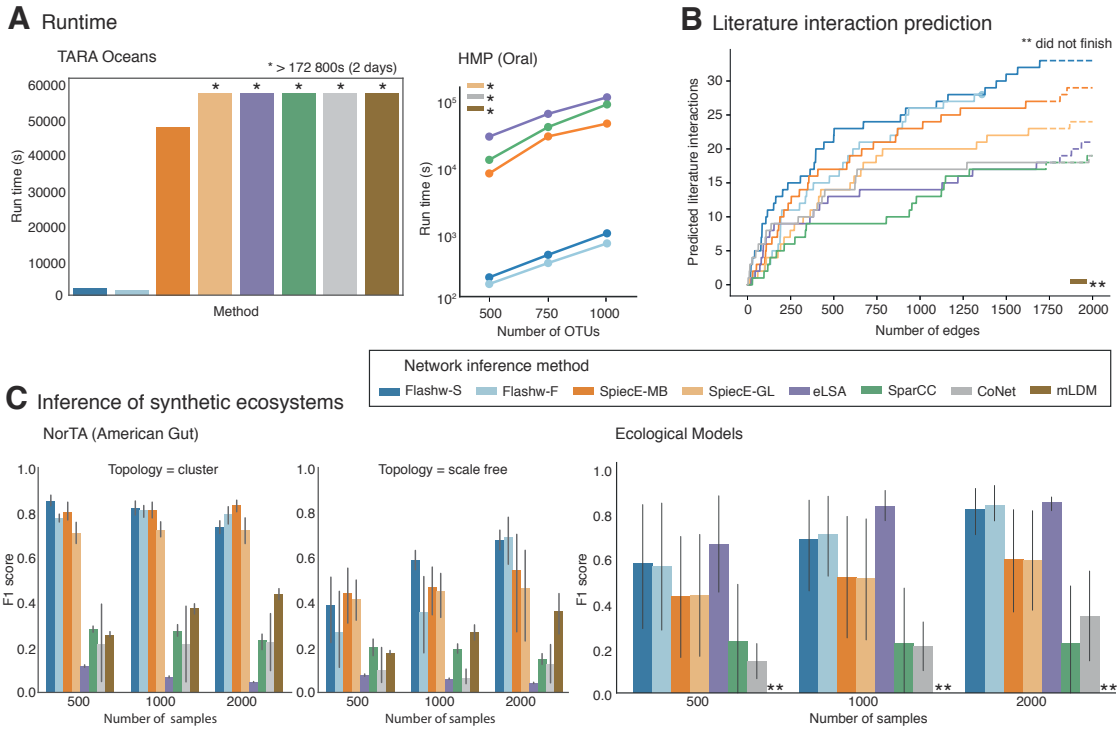


Figure 2: Comparison of FlashWeave to state-of-the-art network inference methods. Method abbreviations are Flashw-S: FlashWeave-S, Flashw-F: FlashWeave-F, SpiecE-MB: SpiecEasi-MB, SpiecE-GL: SpiecEasi-GL. **A** Runtime comparison on the TARA Oceans and Human Microbiome Project (oral body site only) data sets. **B** Number of gold-standard planktonic interactions in the TARA Oceans data set among the 2000 edges ranked most highly by each tool. mLDM did not finish computation after two weeks. **C** Prediction performance on data sets with synthetically engineered edges. Data was generated based on (Kurtz et al. 2015) and (Weiss et al. 2016) and performance measured as F1 score (harmonic mean of precision and recall). Error bars depict 95% confidence intervals of the mean, based on 1000 bootstrap replicates.

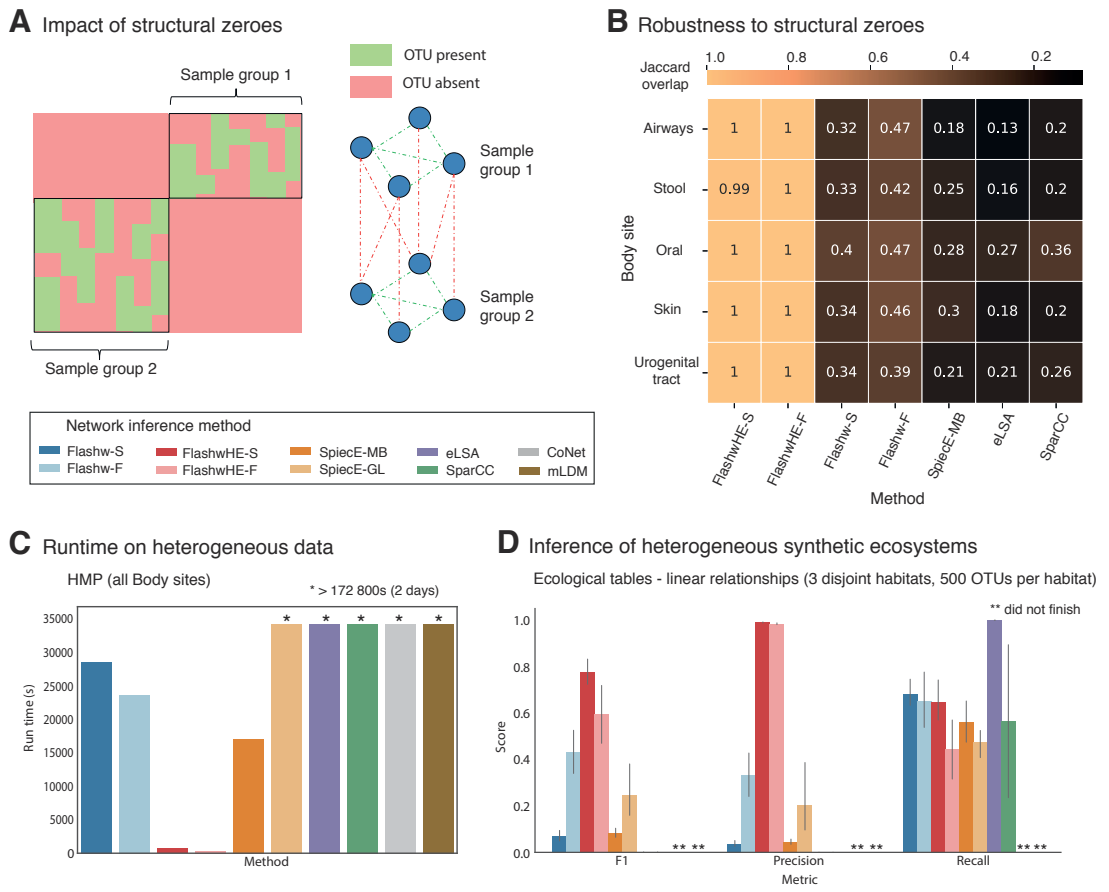


Figure 3: Network inference performance on heterogeneous data sets with increased fractions of structural zeros. Method abbreviations (in addition to those from Fig. 2) are FlashwHE-S: FlashWeaveHE-S, FlashwHE-F: FlashWeaveHE-F. **A** Overview of how structural zeros can lead to false positive edges. **B** Overlap between HMP sub-networks computed on i) OTUs from a single body site (no structural zeros) or ii) body-site specific OTUs from all sites combined (structural zeros). SpiecEasi-GL, CoNet and mLDM did not finish computation after two weeks. **C** Runtime comparison on the HMP data set (all body sites) as a representative heterogeneous data set. **D** Prediction performance on aggregated disjoint habitats generated by the Ecological Models approach (Weiss et al. 2016), measured using F1 score, Recall and Precision. CoNet and mLDM did not finish computation after two weeks. Error bars depict 95% confidence intervals of the mean, based on 1000 bootstrap replicates.

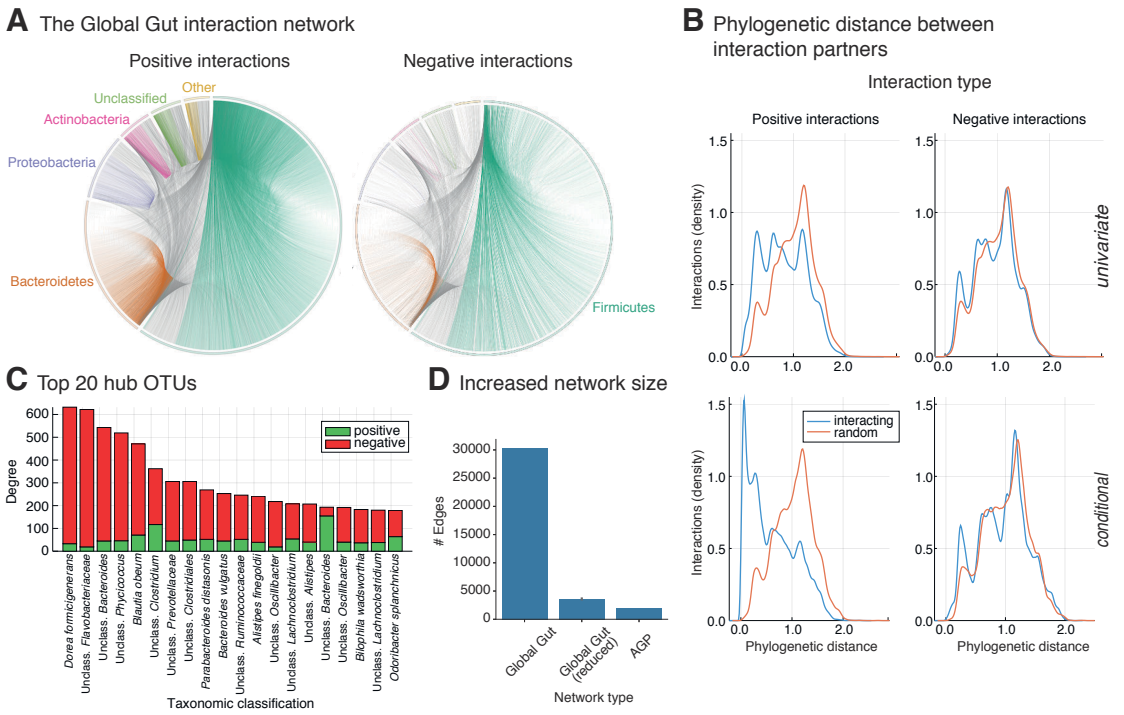
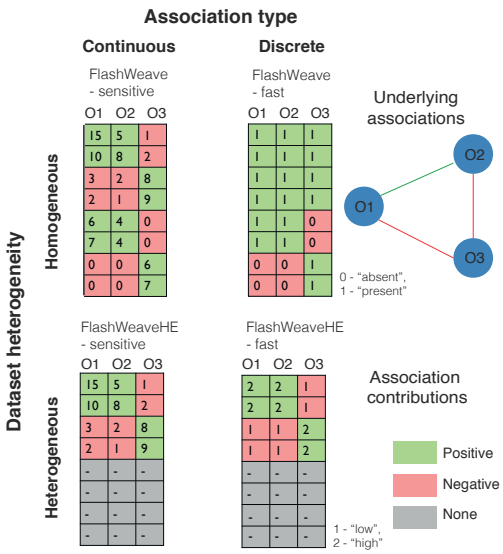
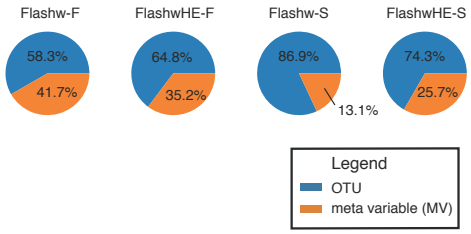


Figure 4: Inference of a large-scale, globally distributed human gut interaction network. **A** High-level overview of positive and negative interactions in the Global Gut network (GGNcond) with OTUs grouped by phylum. Interactions within the same phylum bear that phylum's color, between-phylum edges are shown in grey. **B** Phylogenetic assortativity patterns for positive interactions and negative interactions. Phylogenetic distance is the summed branch length between interaction partners on a tree of 92'659 OTU representatives (98% 16S rRNA identity). The top panel depicts distributions from the univariate (GGNuni) and the lower panel from the conditional Global Gut network (GGNcond). **C** Top 20 OTUs with the highest number of direct interaction partners. OTUs labelled "Unclass." were not confidently classifiable at species level. **D** Comparison of the number of edges between GGNcond, networks of 5 random sample subsets of the GG data set with size equal to the American Gut Project subset, and the American Gut Project network.

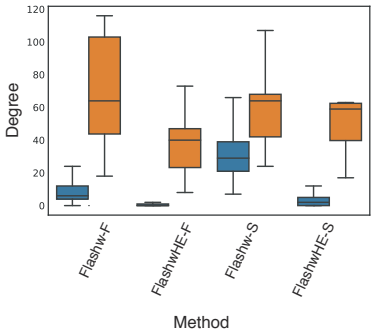
A Differences in statistical information used by each FlashWeave mode



B Fractions of explained associations for OTUs and MVs



C Neighborhood size comparison between OTUs and MVs



D Examples of direct neighborhood reductions of MVs in the HMP

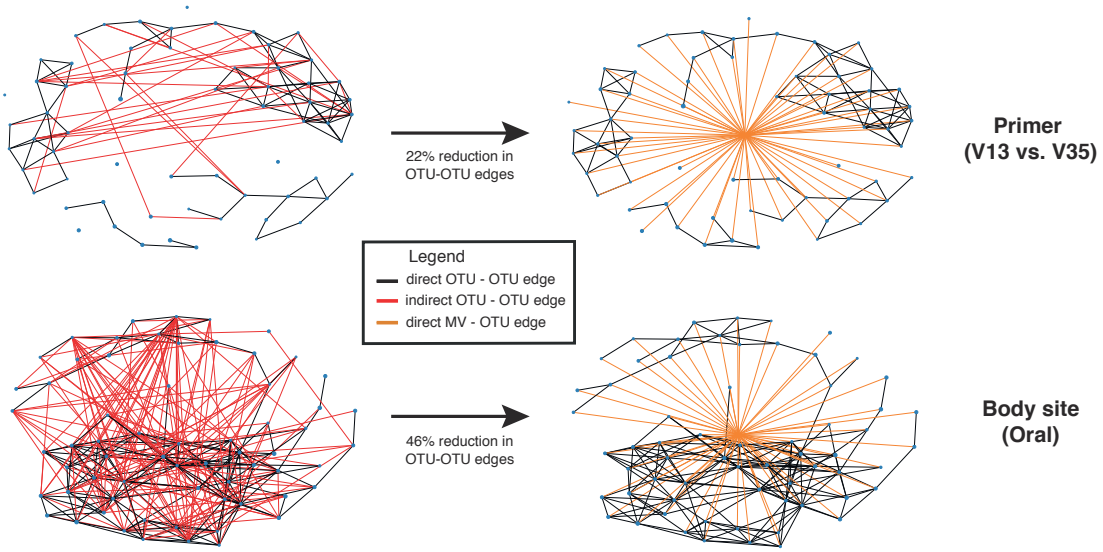
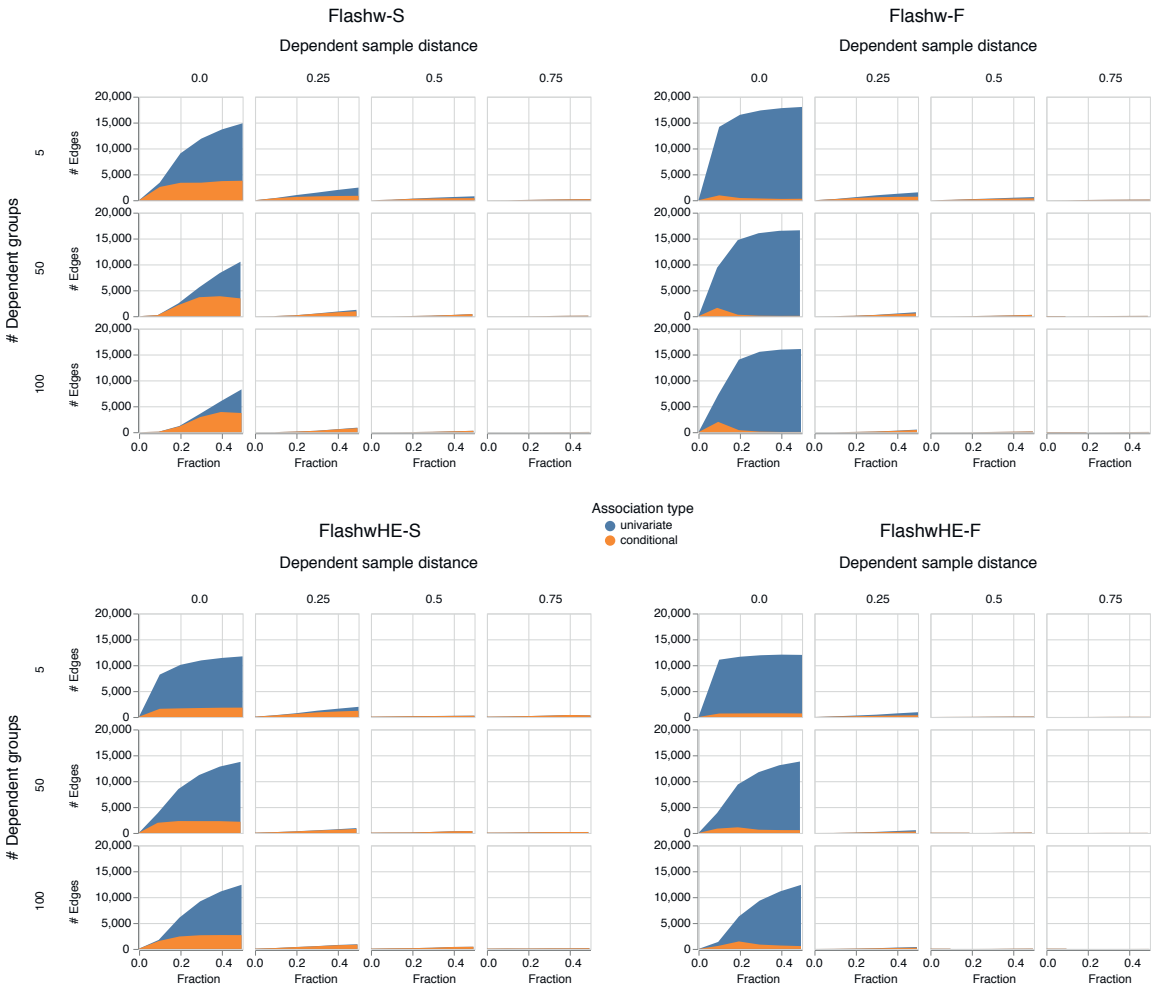


Figure S1: Differences in association information used for each FlashWeave mode and importance of meta variables (MVs) for network inference. **A** Sensitive modes (-S) of FlashWeave use full abundance information ("continuous"), while fast modes (-F) work on discretized abundances. In contrast to FlashWeave, FlashWeaveHE excludes samples in which one partner is absent (colored grey). However, it still includes absences of OTUs within the conditioning sets. **B** Fractions of indirect OTU-OTU associations excluded through at least one MV (orange) or exclusively OTUs (blue) in the HMP network. **C** Number of direct neighbors of OTUs and MVs in he HMP network. **D** Exclusion of indirect edges (red) in the direct neighborhoods of the MVs "Primer" and "Oral" (orange edges) in the HMP network when explicit MV information is provided to FlashWeave-S. Edges retained after including MV information are colored black.

A Influence of simulated dependent sample groups on false positive edge predictions



B Pairwise Bray-Curtis distance within simulated dependent sample groups

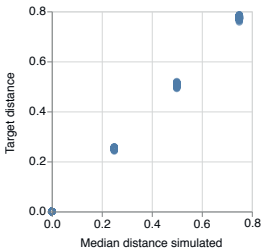
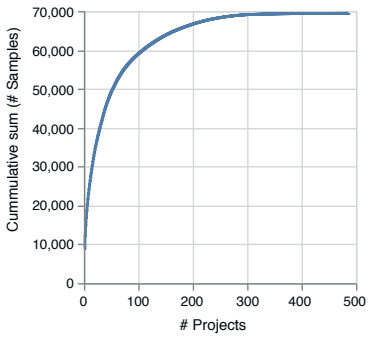
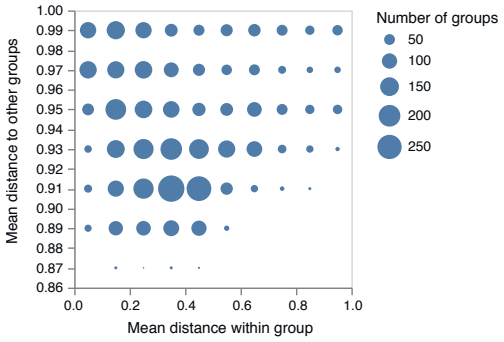


Figure S2: Influence of sample dependencies on network inference. A Simulations of 10'000 samples and 200 OTUs with increasing fractions of dependent samples, increasing numbers of dependent sample groups and increasing Bray-Curtis dissimilarity between dependent samples within the same group. "# Edges" is the number of predicted edges (false positives) for each FlashWeave mode and simulated data set. **B** Pairwise Bray-Curtis dissimilarities between dependent samples in the same group compared to the target distance used as simulation input.

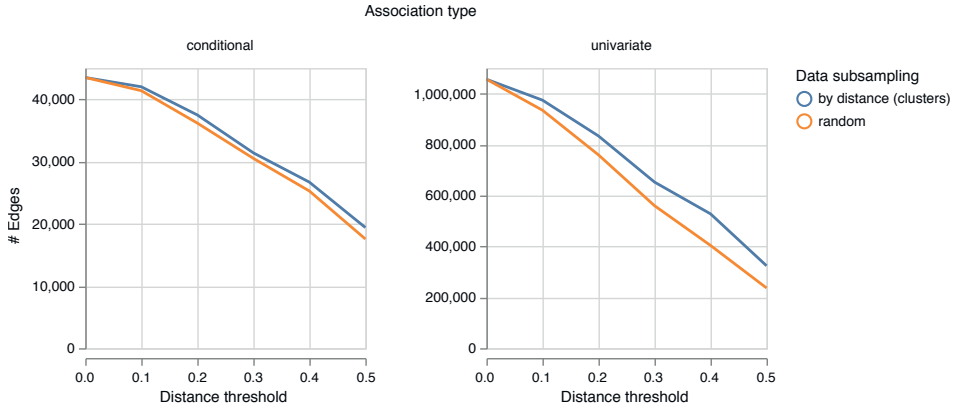
A Number of samples in relation to number of included projects



B Distances within vs. between groups of samples sequenced multiple times



C Impact of sample clustering on the number of network edges



D Top 20 positive hub OTUs

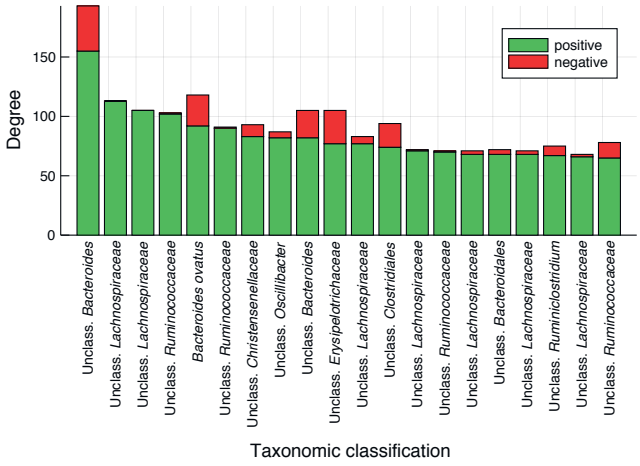


Figure S3: Additional descriptive statistics of the Global Gut data set (GG) and networks (GGNcond, GGNuni). A Increase in the cumulative number of samples in GG when including increasing numbers of projects (ranked by number samples per project, starting with the largest project). **B** For each group of independent sequencing runs of the same sample, depicts the mean distance (Bray-Curtis dissimilarity) between runs within the same group vs. the mean distance to other groups. **C** For successive sample clusterings with increasing Bray-Curtis dissimilarity thresholds, shows the number of edges of inferred networks at each threshold, compared to networks computed from random subsampling of the GG data set with identical sample numbers. **D** Top 20 OTUs with the highest number of direct positive interaction partners. OTUs labelled "Unclass." were not confidently classifiable at species level.

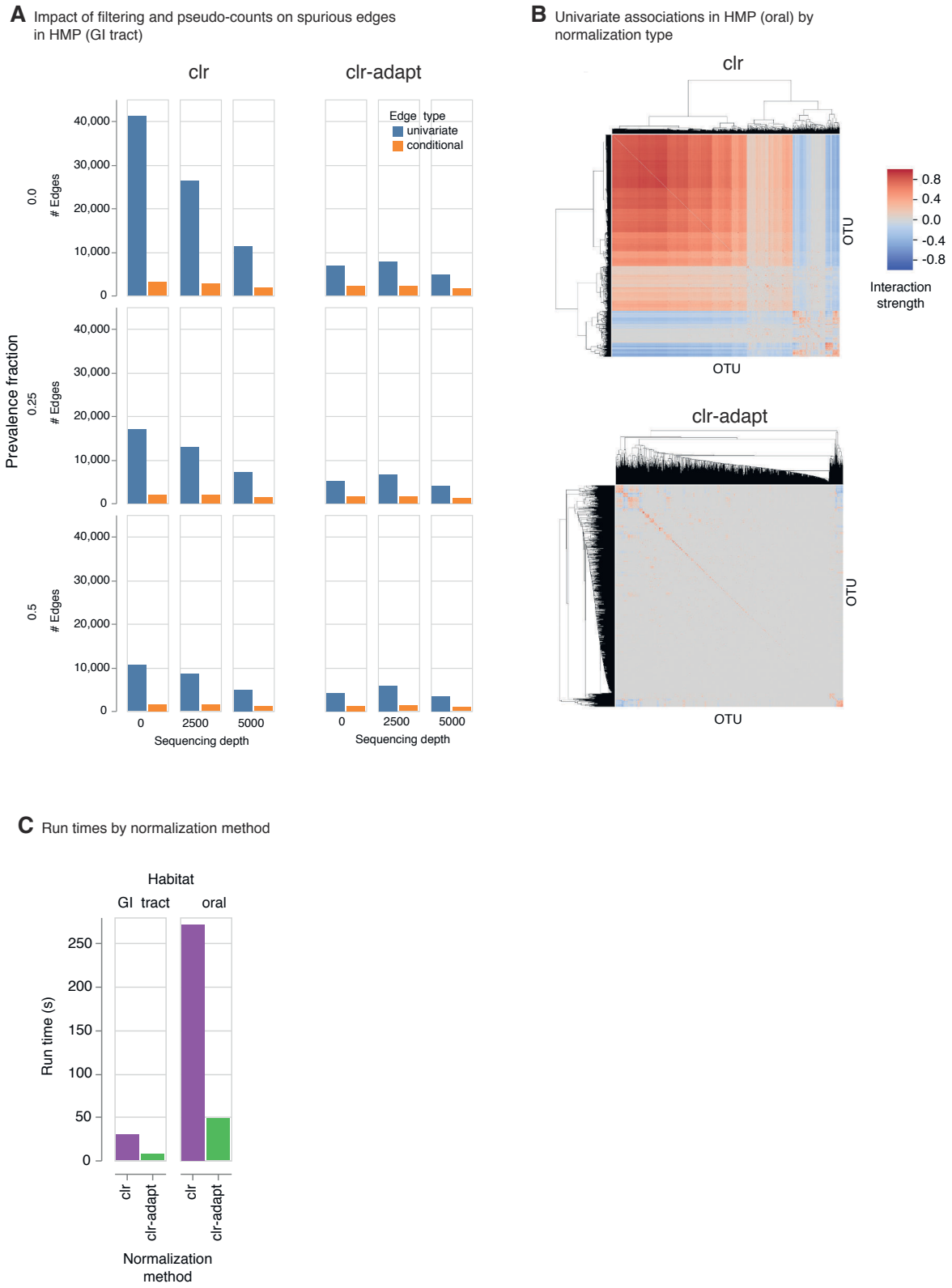

A Impact of filtering and pseudo-counts on spurious edges in HMP (GI tract)
clr
clr-adapt
Edge type
univariate
conditional
Prevalence fraction
0.0
0.25
0.5
Edges
40,000
30,000
20,000
10,000
0
0
2500
5000
Sequencing depth
B Univariate associations in HMP (oral) by normalization type
clr
clr-adapt
OTU
0.8
0.4
0.0
-0.4
-0.8
Interaction strength
C Run times by normalization method
Habitat
GI tract
oral
Run time (s)
250
200
150
100
50
0
clr
clr-adapt
Normalization method

Figure S4: Impact of normalization on network inference. **A** Network sizes along gradients of increasingly strict OTU prevalence (rows) and sequencing depth (columns) filtering of the Gastrointestinal tract subset of the HMP data set, stratified by normalization type (*clr* with pseudo-count 1 vs. *clr-adapt*). **B** Univariate association patterns in the HMP oral network (1000 OTUs) stratified by normalization type (no additional filtering). **C** Runtime comparison between normalization types on HMP (GI tract) and HMP (oral).

Supplemental Text S1

Novel heuristics

Learning the direct neighborhood of target variable T with the si-HITON-PC algorithm has a run time complexity of $O(|V|2^{|PC(T)|})$ (Aliferis et al. 2010), where V are variables in the system and $PC(T)$ is the Parent-Children set of T, i.e. the set of its directly associated neighbors (or the Markov blanket $MB(T)$ minus spouses (Aliferis et al. 2010)). Runtime thus depends linearly on the number of variables and exponentially on the size of the direct neighborhood.

FlashWeave implements all options and algorithmic shortcuts suggested by Aliferis et al. (Aliferis et al. 2010) (max-k heuristic, h-ps reliability criterion, FDR correction, optimal variable ordering). In order to achieve the speed reported in this study, we furthermore extended the original algorithm through a number of additional high-performance heuristics, explained in more detail below.

The first novel algorithmic shortcut we term *feed-forward* heuristic, which is a parallel variation of traditional backtracking (Scutari 2017). The key observation utilized by this heuristic is that the size of individual neighborhoods can vary substantially in networks with scale-free node degree distributions, such as microbial co-occurrence networks ((Faust & Raes 2012) and citations therein), with exponential impact on runtime (see above). From the scale-free property follows that keystone species A (with many dependent neighbors) will typically have a large number of neighbors B that themselves are not keystone species (few neighbors) and whose neighborhoods are thus exponentially quicker to compute. Now, for the edge $A \rightarrow B$ to be included in the final, global network through the OR combinator rule, it is sufficient if the considerably cheaper reverse link $B \rightarrow A$ is proven. *feed-forward* exploits this property by prioritizing computation of variables expected to have smaller neighborhoods (as approximated by their univariate neighborhood size) and relaying the information of detected direct links to computationally more intensive variables (larger neighborhoods). If during the computation of the neighborhood of A the next variable B to be tested was already shown to be a neighbor, it automatically enters the set of neighbor candidates of A without formally performing all tests. *feed-forward* is applied in a parallel computation setting, where candidate lists of all nodes are periodically updated with the latest information from other neighborhoods as it becomes available, allowing the most expensive nodes to leverage a maximum amount of information to cut down runtime.

The second computational shortcut introduced in FlashWeave we term *fast-elimination* heuristic. A large amount of time can be spend in the final elimination phase of si-HITON-PC, in which all previously skipped tests between candidates passing the interleaving phase are performed. In the original algorithm, even if a variable S is discarded during the elimination phase, it will still be included in future conditioning sets, thereby inflating the number of conducted conditional independence tests. If many variables are discarded during earlier parts of the elimination phase, but still included in subsequent conditioning sets, the result can be an exponential

increase in necessary tests, making the elimination phase particularly costly. *fast-elimination* addresses this computational hurdle by not considering a removed variable S for any subsequent conditioning sets during elimination phase. An intuitive motivation for this approach is that, if a variable S was shown earlier to not be part of the neighborhood of T, it should also not be required to render further candidates independent of T as it's not part of its Parent-Children set.
As another shortcut, we implemented a convergence criterion that periodically checks whether links in the network still show substantial change over time. If the network has reached convergence, all remaining candidates are assumed to be conditionally independent of their target variables. This criterion is based on our observation that the naive algorithm can stall on single nodes with large neighborhoods due to the exponential runtime dependency of si-HITON-PC on neighborhood size. However, candidates still to be checked at this point tend to be weak, since they i) appear late in the relevance-sorted candidate list and ii) have been proven to not be neighbors in the reverse direction (otherwise the *feed-forward* heuristic would have applied). The majority of these late links are typically discarded after substantial computational effort, with minimal effect on network structure. While using this type of convergence threshold may in theory lead to biased edge omissions since it selectively bypasses computation of candidates in large neighborhoods, we didn't detect meaningful biases of this kind in the networks we tested.
As a final option to improve runtime, FlashWeave can be instructed to run only up to a certain (by default large) number of tests per node, assuming that performing such a high number of tests provides reasonable safety that the current candidate will not be discarded by additional tests. This effectively puts an upper bound on the exponential behaviour of si-HITON-PC and helps to prevent extensive run times on single variables with large neighborhoods, with empirically minimal effect on network structure. However, FlashWeave will flag these interactions and warn the user in case the boundary is breached.

Normalization

Sequencing data is subject to mainly technically determined and thus arbitrary variations in sequencing depth, making it compositional in nature (Papageorgiou & Aitchison 1989; Pawlowsky-Glahn & Buccianti 2011). Compositionality impedes naive correlation analysis without adequate correction (Aitchison 1981; Friedman & Alm 2012). Common approaches to properly analyze compositional data include various log-ratio based methods, such as log-ratio transformations (Aitchison 1981).
Similar to SpiecEasi (Kurtz et al. 2015), FlashWeave uses the centered log-ratio (*clr* (Aitchison 1981)) approach for compositionality correction of a vector x of compositional microbial abundances:

$$clr(x_{ij}) = log \frac{x_{ij}}{g(s_i)} \quad \text{with } g(s_i) = \left[\prod_{l=1}^{p} x_{il} \right]^{\frac{1}{p}} \tag{1}$$

where $g(s_i)$ describes the geometric mean of all compositional abundances in sample s_i, p the total number of OTUs and $clr(x_{ij})$ the clr-transformed value of the compositional abundance of microbe j in sample s_i.

An inherent shortcoming of logarithm-based methods is the handling of absences (zeroes) in the input data. This is usually circumvented by applying a fixed pseudocount (for example 1) to the input data which then allows proper logarithmic computations. Our analyses revealed that this approach can work for strongly filtered and depth-homogeneous data sets, but introduces noticeable biases when applied to data sets including rare OTUs and samples with particularly low sequencing depths (Fig. S4 A, left column). In such data, we observed extensive increases in univariate network density, which rendered the subsequent conditioning search in FlashWeave unusually slow (Fig. S4 C). Importantly, most of these additional univariate associations are finally removed during conditioning search (Fig. S4 A, left column), indicating their spurious nature.

More precisely, absences of comparatively rare OTUs in low-depth samples can, after clr transformation, become values higher than the OTU's mean clr-transformed abundance across all samples, while absences in high-depth samples result in transformed values below these OTU's means. This depth-based deviation from the mean results in the observed artificial association signal and notably is driven solely by applying the same fixed pseudo-count both to low-depth and high-depth samples. While homogenizing sequencing depth through sample removal and filtering of rare OTUs reduces this signal (Fig. S4 A, left column), large amounts of valuable data are potentially removed by this approach.

As an alternative method to reduce the pseudo-count driven association signal, we suggest a modification to classic fixed pseudo-counts, which we term "adaptive pseudo-counts", resulting in the normalization scheme ***clr-adapt***. In this approach, initially a fixed pseudo-count π_{max} is applied to the sample with the highest sequencing depth (s_{max}). Then solving

$$log(\frac{\pi_i}{g(s_i)}) = log(\frac{\pi_{max}}{g(s_{max})}) \tag{2}$$

for π_i (the adaptive pseudocount for sample s_i) leads to

$$\pi_i = \left[\frac{\pi_{max}^{k-p} \cdot g_{nz}(s_{max})}{g_{nz}(s_i)} \right]^{\frac{1}{n-p}} \tag{3}$$

where $g_{nz}(s)$ is the geometric mean of all non-zero abundances in sample s, k is the number of absences in sample s_{max} and p is the number of OTUs. Formula 3 is applied to all samples excluding s_{max} in order to determine sample-specific adaptive pseudo-counts. These are then applied to their respective samples, followed by usual clr transformation (formula 1). This results in the same transformed absence counts in all samples and ensures that absences are below each OTU's mean clr-abundance, which avoids bi-directional pseudo-count driven deviations from the mean. Using this approach, we observe strongly reduced univariate network densities, discard fractions and run times (Fig. S4).

FlashWeaveHE also utilizes clr transformation for compositionality correction, albeit slightly modified. Since FlashWeaveHE does not consider absences for its association calculations (see Methods), it requires no (adaptive) pseudo-counts. Instead, only non-zero compositional abundances are used to compute the compositional center (geometric mean, formula 1) used for the transformation, resulting in the normalization scheme *clr-nonzero*.

FlashWeave-F and FlashWeaveHE-F differ from the FlashWeave-S and FlashWeaveHE-S by applying mutual information tests which necessitate data discretization. FlashWeave-F uses a straight-forward discretization scheme: all non-zero abundance values become one, while absences remain zero. This approach makes clr normalization and pseudo-counts unnecessary and is inherently robust to compositional artifacts. It is furthermore less affected by sequencing depth differences than raw numbers of reads, since depth only affects OTU-presence and absence but not abundance. FlashWeaveHE-F on the other hand discretizes all *clr-nonzero* transformed values into two bins per OTU ("high" abundance vs "low" abundance), with bins separated by the median.

Meta variables (MVs) are by default not normalized for FlashWeave-S and FlashWeaveHE-S and should thus, if necessary, be provided in a sensible pre-normalized format by the user. For FlashWeave-F and FlashWeaveHE-F, continuous MVs are by default discretized into two bins separated by their median.

Estimation of dependent sample group influence on the Global Gut network

In order to estimate the impact of dependent sample groups on network inference of the Global Gut data set (GG), we first identified samples that were sequenced more than once. This resulted in 4700 samples (9% of all samples), independent re-sequencings of which covered 31% of the data set in total. When analyzing sample distances (measured as Bray-Curtis dissimilarity) within vs. between these sample groups, we found that within-group distances, while generally smaller than between-group distances, still covered a wide range of values (Fig. S3 B). This indicated substantial variation within sample groups, potentially providing important information for network inference. Additionally, our simulation benchmarks predicted negligible numbers of false positives for the mean within-group distance (0.37), dependent sample fraction (31%), number of groups (> 100) and mode (FlashWeaveHE-F) used to compute the Global Gut

network (GGNcond) (Fig. S2 A). As expected from our simulations, we furthermore detected a steep increase in predicted edges (43'493 to 82'552, an increase of 89%) when replacing each group with identical copies of one group-specific representative, and thus shifting the data set towards the distance region with high expected numbers of false positives.
Despite shared sample material, other sources of sample dependence, such as samples taken from the same individual in a time series, could also influence GGNcond. To account for these types of dependence, we systematically clustered samples in GG with increasing sample distance thresholds. For each clustering, we then computed networks based on only cluster representatives and compared these to a background of networks computed from random subsets of GG with matching numbers of samples. In the presence of substantial false positives due to sample dependence, we expected the cluster-based networks to show steeper initial drops in edge numbers compared to the background networks, because clustering specifically removes spurious dependence signals and hence false positives, while random subsampling retains them. In addition, univariate networks produced markedly increased numbers of false positives under strong sample dependence conditions than conditional networks in our simulations (Fig. S2 A), suggesting that initial drops in edge numbers should be even more distinct for univariate networks. In our tests, we however did not observe any of these clear indicators of sample dependence, since edge reductions in cluster-based networks were always similar to or smaller than for background networks, both in the conditional and univariate case (Fig. S3 C). Hence, we could not detect a signal for false positives due to sample dependence in GGNcond.

Supplemental references

Aitchison, J., 1981. A new approach to null correlations of proportions. *Journal of the International Association for Mathematical Geology*, 13(2), pp.175–189.

Aliferis, C.F. et al., 2010. Local causal and markov blanket induction for causal discovery and feature selection for classification part i: Algorithms and empirical evaluation. *Journal of machine learning research: JMLR*, 11(1), pp.171–234.

Faust, K. & Raes, J., 2012. Microbial interactions: from networks to models. *Nature reviews. Microbiology*, 10(8), pp.538–550.

Friedman, J. & Alm, E.J., 2012. Inferring correlation networks from genomic survey data. *PLoS computational biology*, 8(9), p.e1002687.

Kurtz, Z.D. et al., 2015. Sparse and compositionally robust inference of microbial ecological networks. *PLoS computational biology*, 11(5), p.e1004226.

Papageorgiou, H. & Aitchison, J., 1989. The Statistical Analysis of Compositional Data. *Biometrics*, 45(1), p.345.

Pawlowsky-Glahn, V. & Buccianti, A., 2011. *Compositional Data Analysis: Theory and Applications*, John Wiley & Sons.

Scutari, M., 2017. Bayesian Network Constraint-Based Structure Learning Algorithms: Parallel and Optimized Implementations in the bnlearn R Package. *Journal of statistical software*, 77(2). Available at: http://dx.doi.org/10.18637/jss.v077.i02.

2.2 Ecologically informed microbial biomarkers and accurate classification of mixed and unmixed samples in an extensive cross-study of human body sites

Contribution: Janko Tackmann (JT) contributed to project conceptualization, as well as data collection and preprocessing. JT furthermore created the pipelines for classifier training, implemented the software for biomarker detection, performed all benchmarks and made major contributions to the interpretation of results. In addition, JT generated the majority of visualizations found in the manuscript, made substantial contributions to the initial manuscript and contributed to reviewing and editing of the final manuscript.

Tackmann et al. Microbiome (2018) 6:192
https://doi.org/10.1186/s40168-018-0565-6

Microbiome

RESEARCH Open Access

Ecologically informed microbial biomarkers and accurate classification of mixed and unmixed samples in an extensive cross-study of human body sites

Janko Tackmann[1], Natasha Arora[2], Thomas Sebastian Benedikt Schmidt[1,3], João Frederico Matias Rodrigues[1†] and Christian von Mering[1*†]

Abstract

Background: The identification of body site-specific microbial biomarkers and their use for classification tasks have promising applications in medicine, microbial ecology, and forensics. Previous studies have characterized site-specific microbiota and shown that sample origin can be accurately predicted by microbial content. However, these studies were usually restricted to single datasets with consistent experimental methods and conditions, as well as comparatively small sample numbers. The effects of study-specific biases and statistical power on classification performance and biomarker identification thus remain poorly understood. Furthermore, reliable detection in mixtures of different body sites or with noise from environmental contamination has rarely been investigated thus far. Finally, the impact of ecological associations between microbes on biomarker discovery was usually not considered in previous work.

Results: Here we present the analysis of one of the largest cross-study sequencing datasets of microbial communities from human body sites (15,082 samples from 57 publicly available studies). We show that training a Random Forest Classifier on this aggregated dataset increases prediction performance for body sites by 35% compared to a single-study classifier. Using simulated datasets, we further demonstrate that the source of different microbial contributions in mixtures of different body sites or with soil can be detected starting at 1% of the total microbial community. We apply a biomarker selection method that excludes indirect environmental associations driven by microbe-microbe associations, yielding a parsimonious set of highly predictive taxa including novel biomarkers and excluding many previously reported taxa. We find a considerable fraction of unclassified biomarkers ("microbial dark matter") and observe that negatively associated taxa have a surprisingly high impact on classification performance. We further detect a significant enrichment of rod-shaped, motile, and sporulating taxa for feces biomarkers, consistent with a highly competitive environment.

Conclusions: Our machine learning model shows strong body site classification performance, both in single-source samples and mixtures, making it promising for tasks requiring high accuracy, such as forensic applications. We report a core set of ecologically informed biomarkers, inferred across a wide range of experimental protocols and conditions, providing the most concise, general, and least biased overview of body site-associated microbes to date.

Keywords: Human microbiome, Biomarkers, Mixture, Random Forest, Generalized Local Learning

* Correspondence: mering@imls.uzh.ch
†João Frederico Matias Rodrigues and Christian von Mering contributed equally to this work.
[1]Institute of Molecular Life Sciences and Swiss Institute of Bioinformatics, University of Zurich, Zurich, Switzerland
Full list of author information is available at the end of the article

Background

The identification of microbial biomarkers and their use for classification provide valuable information for predictions in a wide range of applications. For example, disease states can be associated with complex microbial patterns, making suitable biomarker detection and classification methods crucial for accurate disease identification and prediction [1–4]. Another important application is the identification of the source of origin for a sample, for example in forensic cases [5] or environmental monitoring [6]. In the former, reliably determining the bodily source of a stain at a crime scene (e.g., saliva, semen, vaginal fluid, blood) can critically aid the reconstruction of crime events. In the latter, distinguishing microbial communities of human origin from those of environmental provenance can be of great value, for example in studies of human sewage pollution [7]. Establishing the source of microbial communities has an added complexity when dealing with mixtures, such as in contaminated samples, as these require the distinction of two or more different sources.

In the last decade, advances in high-throughput sequencing have led to a marked increase in both the number and sequencing depth of human microbiome studies, encompassing a wide range of experimental protocols and conditions (e.g., sex, geography, medication) [8, 9]. This increase in data volume and diversity has opened the door for human microbiome studies to apply more advanced statistical and machine learning tools [10, 11].

Pioneering studies have explored the potential of supervised machine learning approaches on a number of microbiome-based classification tasks, such as identification of individual subjects, disease prediction, and body site identification [10–12], mostly focusing on comparatively small-scale datasets from individual studies [13, 14]. In these evaluations, the best-performing methods generally achieved high predictive power, with Random Forest Classifiers (RFCs, [15]) ranking consistently among the top performing models. Further studies have successfully applied classification methods to varying sample types such as human, soil, and sediments [16–18] and also in the context of multi-source mixtures [19, 20].

While the success of supervised learning methods has been demonstrated in individual studies, it is largely unclear whether these results generalize to a meta-study setting. For instance, a large fraction of publicly available data consists of amplicon sequencing runs, where primer specificity can result in amplification biases [21, 22]. This can affect taxonomic inferences and classification accuracy, potentially leading to classifiers with top performance for some primer types and mediocre performance for others. While metagenomic shotgun sequencing tends to produce less biased estimates of community composition, this approach is more expensive and typically results in a lower coverage of the 16S rRNA gene, potentially compromising the resolution of abundances for rare taxa [23]. Apart from protocol-specific effects, subject-specific factors such as geographic location and medication may introduce further biases. Recent work in disease prediction also suggests that individual studies may report unspecific signals only properly appreciated when aggregating studies [24].

The aggregation of sequencing data from different studies into large meta-datasets and their utilization for classifier training could thus lead to more general and predictive models that can reliably classify samples produced under a variety of experimental protocols and from a wide range of subjects. It further may allow identification of biomarkers that overcome biases of individual studies. Recent work by Pasolli et al. [25] showed promising results for the predictive power of cross-study models for classification of diseased versus healthy subjects, as well as classification of body sites. However, their body site classification analysis was based on 642 sequencing samples from only two studies and furthermore restricted to whole genome shotgun sequencing. Therefore, it is yet unclear whether these results generalize across larger and experimentally more heterogeneous datasets.

Such datasets are of particular relevance for the identification of microbes endemic to human body sites. Site-specific microbes have been extensively studied in large cooperative efforts, such as the Human Microbiome Project (HMP, [26]), and in smaller studies by individual groups. However, single-study biases and insufficient sample size could significantly influence previously reported associations. For instance, the original HMP was restricted to 242 healthy adults situated in the USA and limited to two primer sets, raising questions about whether reported site-specific taxa generalize across geography, subject-specific conditions, and experimental protocols.

While some microbes are endemic to a body site, others are only indirectly associated with a site due to their ecological dependency on endemic microbes. These indirectly associated microbes could in principle also thrive in other habitats, where the same requirements may be fulfilled by other partners. Common biomarker identification approaches may misinterpret this ecological signal and specify such indirectly associated microbes as body site biomarkers. More refined methods, on the other hand, can separate directly and indirectly associated markers by testing whether an association signal can be explained by other variables [27, 28]. Generalized Local Learning (GLL, [27]), a method that excludes indirectly associated biomarkers, was previously shown to achieve the best balance

between number of identified biomarkers and accuracy in psoriasis prediction [12]. Importantly, such ecological interaction effects were usually not considered by previous studies on biomarkers for human body sites, potentially inflating numbers of reported markers.

In this study, we had two aims: (i) to train a supervised classification model on a heterogeneous, large-scale dataset and evaluate its performance compared to a single-study classifier in body site prediction, both for single-source samples and mixtures, and (ii) to obtain a high-quality set of microbial biomarkers directly associated with human body sites, excluding microbe-microbe-driven associations. To this end, we retrieved 50,273 publicly available sequenced samples from five human body sites (skin, saliva, vagina, nostril, and feces), further filtered down to 15,082 classification-ready samples spanning 57 studies. We additionally retrieved sequencing data for 1329 soil samples as representatives of a typical environmental contaminant. We used the body site dataset for classifier training and evaluated its performance on single-source samples, as well as in silico mixtures of samples from two body sites or from a body site and soil. We compared this performance to a classifier trained on a single study, subject to previous machine learning benchmarks, and demonstrated that the cross-study classifier makes strongly improved predictions. Finally, we identified a parsimonious core set of microbial biomarkers for the investigated body sites, which included previously unreported biomarkers and, mostly due to our bias-mitigating cross-study approach and the exclusion of indirect associations, rejected previously reported study-specific associations. We analyzed this set of biomarkers in depth in terms of taxonomy, phylogeny, and physiological traits.

Results

A large and heterogeneous collection of microbial sequencing samples from human body sites

Publicly available metagenomic sequence read data were retrieved from the NCBI Sequence Read Archive [29] for studies investigating microbial communities in five human body sites: nostril, saliva, skin, vagina, and feces. The initial dataset consisted of 50,273 samples, sequenced mainly through targeted amplicon sequencing and whole genome shotgun sequencing technology. After extensive filtering (see the "Methods" section), this dataset was reduced to a condensed set of 15,082 samples (see Additional file 1: Table S4 for accession numbers) from 57 studies, where the number of samples per body site ranged between 1354 (nostril) and 5296 (skin). In total, 60,892 operational taxonomic units (OTUs) were identified after mapping to a database of 16S rRNA reference sequences provided by MAPseq [30], pre-clustered at 96% sequence similarity (Fig. 1). We refer to this dataset as GlobalBodysites.

Cross-study classifier outperforms single-study model in predictive accuracy

We trained an optimized Random Forest Classifier (RFC-global) on the GlobalBodysites dataset and assessed its performance on labelling samples with their correct body sites in a fivefold cross-validation framework (see the "Methods" section). Performance was measured through F1 scores, which take into account both precision and recall and are less affected by imbalanced numbers of samples per body site than other metrics. The classifier was able to accurately identify body site labels in the cross-validated test sets (Fig. 2a), with mean F1 scores between 0.73 (nostril) and 0.95 (feces). Training and testing the classifier on sample sets with biased body site proportions yielded comparable results (Additional file 2: Figure S10b).

We next investigated which pairs of body sites were most challenging to distinguish for RFC-global. To this end, we generated a confusion matrix, capturing the misclassifications across all pairs of body sites (Fig. 2b). We observed a relatively small number (5 to 11%) of misclassifications for all site pairs except one: nostril samples were more prone to misclassification (36%), with a majority of mislabellings as skin (33%). This pattern is in line with nostril-skin misclassifications observed in previous work [25].

Next, we compared the predictive performance of RFC-global to a classifier trained only on the samples from a single study. To this end, we trained a RFC on the subset of GlobalBodysites comprising 372 samples from Costello et al. [13] (RFC-single), a dataset used extensively for body site prediction benchmarks in previous research [10, 11]. We found the prediction performance for RFC-global (trained across studies) to be considerably higher than that for RFC-single (mean F1 of 0.89 compared to 0.66, an increase of 35%) (Fig. 2c).

In order to further elucidate this difference in predictive performance, we looked at the intrinsic feature importance assignments of the RFCs in detail. Briefly, RFCs assign a weight (feature importance) to each OTU depending on its estimated predictive importance. We found that 91.4% of the 1397 predictive OTUs (feature importance > 0) reported by RFC-single were also supported by RFC-global. However, RFC-global reported 15,863 additional predictive OTUs (Additional file 3: Figure S1a,b), an increase by a factor of more than 12. Additionally, 8.6% of the OTUs predictive for RFC-single were dismissed as uninformative by RFC-global. We further observed that

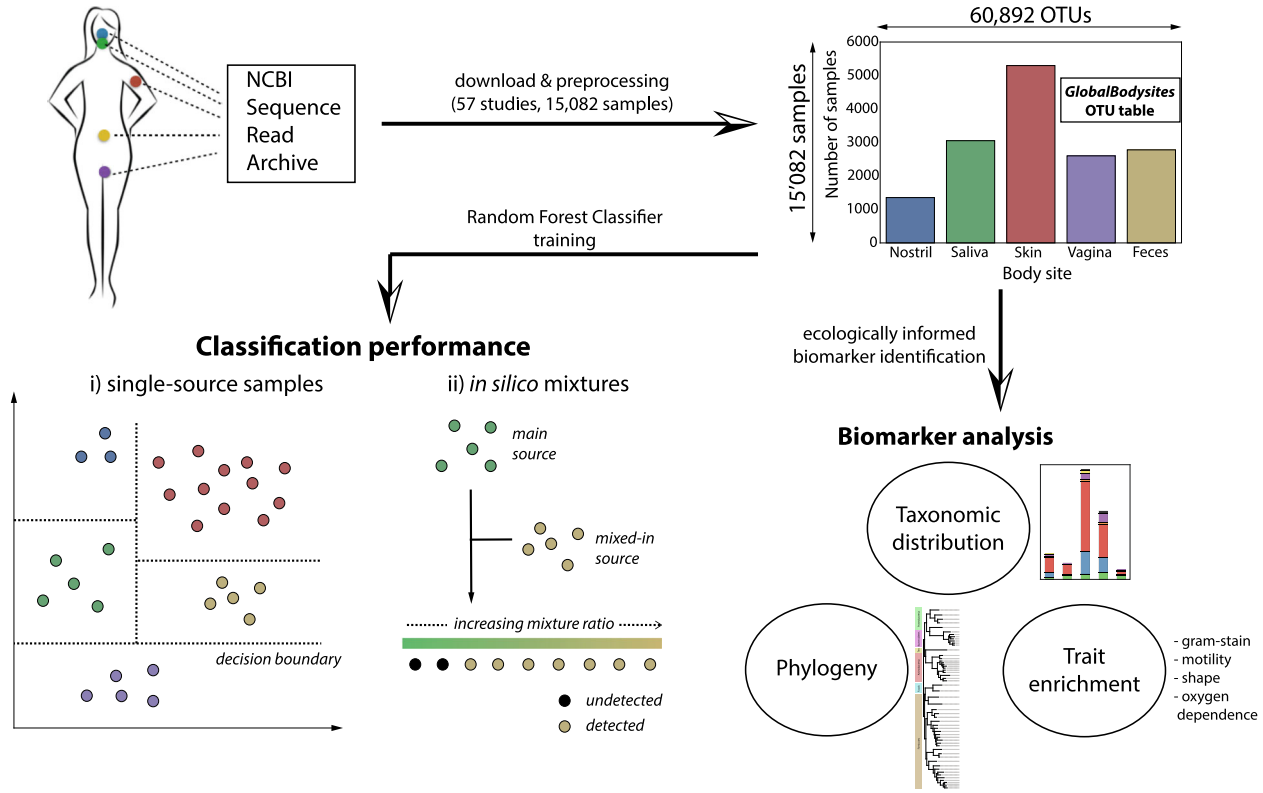


Fig. 1 Fifteen thousand eighty-two samples from five body sites and 57 studies were queried from the NCBI Sequence Read Archive and uniformly processed, yielding the GlobalBodysites dataset. A Random Forest Classifier (RFC-global) was trained and evaluated on this dataset, followed by identification and analysis of ecologically informed microbial biomarkers

feature importances of overlapping predictive OTUs were only moderately correlated (Spearman's rho of 0.68, p value < 0.05, Additional file 3: Figure S1c).

In order to estimate whether increased prediction performance extends to other single studies, we further trained a classifier on all 5433 samples belonging to the Human Microbiome Project (HMP, RFC-single-hmp), which constitutes the largest single study in GlobalBodysites (36% of all samples). While performance of RFC-global was closer to RFC-single-hmp (12% F1 score increase, Additional file 2: Figure S10a, "unweighted"), we found this greater similarity to be driven by the

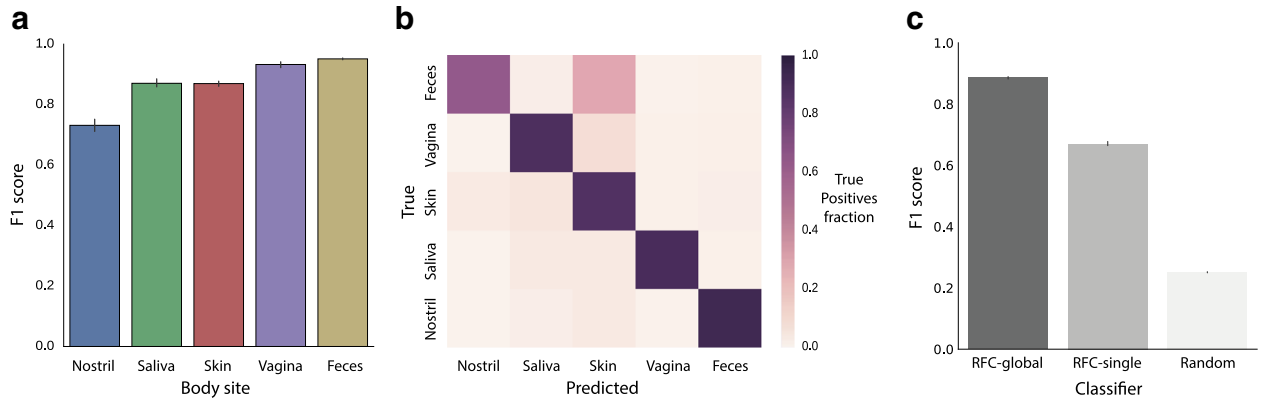


Fig. 2 **a** Prediction performance of RFC-global was estimated on five cross-validation sets of unseen samples, measured as F1 score. **b** Distribution of true positive and false positive classifications across all pairs of body sites. **c** Comparison of RFC-global to a RFC trained on only samples from the Costello et al. dataset (RFC-single) and a random guesser, measured as F1 score

dominant fraction of HMP samples in our validation sets. When weighting studies by the inverse of their sample numbers in F1 score calculation, effectively increasing the reward for correctly predicting samples from smaller studies, we found the performance increase of RFC-global to be noticeably more pronounced (41% F1 score increase, Additional file 2: Figure S10a, "weighted").

We next investigated whether the performance of RFC-global could benefit from additional data. To this end, we tested classification performance for increasing numbers of studies and samples, starting with only the HMP project (Additional file 4: Figure S11). For the "unweighted" scenario (see previous paragraph), we found that performance plateaued around 30 studies (~ 8000 samples), indicating small benefits from additional studies for classifier performance. However, when increasing the reward of predicting samples from smaller studies ("weighted"), we found that performance did not reach a plateau but instead consistently increased with more studies.

Even trace amounts of body site microbiomes can be reliably identified in mixtures between body sites or body site and environment

Next, we evaluated detection limits and prediction performance of RFC-global on in silico mixtures of different body sites (see the "Methods" section). The task was to identify a microbial community from a target body site in mixtures of communities from the target body site and a background body site, along a gradient of increasing mixture fractions, for all pairs of body sites. Classification performance was measured through the area under the ROC curve (AUC) metric, which similarly to F1 scores is robust to label imbalances, but quantifies predictive performance independent of a decision threshold.

At 15% mixture fraction, RFC-global achieved more than 75% of the AUC obtained for unmixed samples in 15 out of 20 body site combinations (Fig. 3). For six of these combinations, these AUC scores were achieved even at trace amounts as low as 1 to 2% mixture fraction. Highest performance was reached when the source body site was vagina or feces, with an average AUC of 0.81 at 2% mixture fraction. In agreement with the confusion matrix of single-source samples (Fig. 2b), distinguishing skin and nostril samples was challenging: mixtures containing these two body sites required at least 70% skin for the AUC to reach 0.8 when skin was the target body site (Additional file 5: Figure S2). Interestingly, the identification of nostril in mixtures containing skin required a mixture fraction of only 30% to achieve the same AUC.

Compared to RFC-global, RFC-single was considerably less robust to mixed sources: predictions resulted in lower AUC values irrespective of which body sites were mixed, in some cases down to little more than random guessing (AUC < 0.6), even when the target body site comprised up to 80% of the mixture (Additional file 6: Figure S4).

Since AUC measures general discriminative power without setting a specific classification threshold, we also determined thresholds above which the presence of a body site fraction in a mixture was likely. To this end, we estimated optimal prediction thresholds for all body site combinations based on the training data and computed F1 scores for each pair of body sites (Additional file 7: Figure S3). For most pairs, F1 results were comparable to the AUC analyses, but some combinations required higher fractions of the target body site when imposing a fixed threshold.

We next investigated the predictive performance and robustness of our classifier in mixtures comprising bacterial communities from a human body site and an environmental component. We prepared in silico mixtures between body site samples from the GlobalBodysites data and 1329 additional microbial soil samples from the NCBI SRA database. For all non-fecal mixtures, we obtained AUC values greater than 0.9 even in samples that consisted mostly of soil (body site mixture fractions below 10%) (Additional file 8: Figure S5). For feces, performance was slightly decreased to between 0.8 and 0.9 AUC for mixture fractions below 50%.

In order to also test the robustness of the classifier to contamination in training samples, we randomly mixed 50% of all training samples with 30% soil, followed by classifier training (RFC-global-contaminated). Validation of RFC-global-contaminated on unmixed body site samples resulted in F1 scores similar to RFC-global (mean decrease 1.3%).

A parsimonious core set of directly associated microbial biomarkers for human body sites

Having assessed the predictive power of RFC-global, we were next interested in biological patterns driving its performance and whether new or unusual associations between microbes and environment could be detected in GlobalBodysites. We thus sought to identify a core set of microbial biomarkers for each investigated body site.

We used Generalized Local Learning (GLL, [27]) (Fig. 4a), an approach that has advantages over feature importances reported by Random Forests and decision trees ([31], see Methods).

Briefly, GLL only reports OTUs as biomarkers whose association with a habitat cannot be statistically explained by ecological dependencies on other OTUs in that habitat. This effectively makes the identified

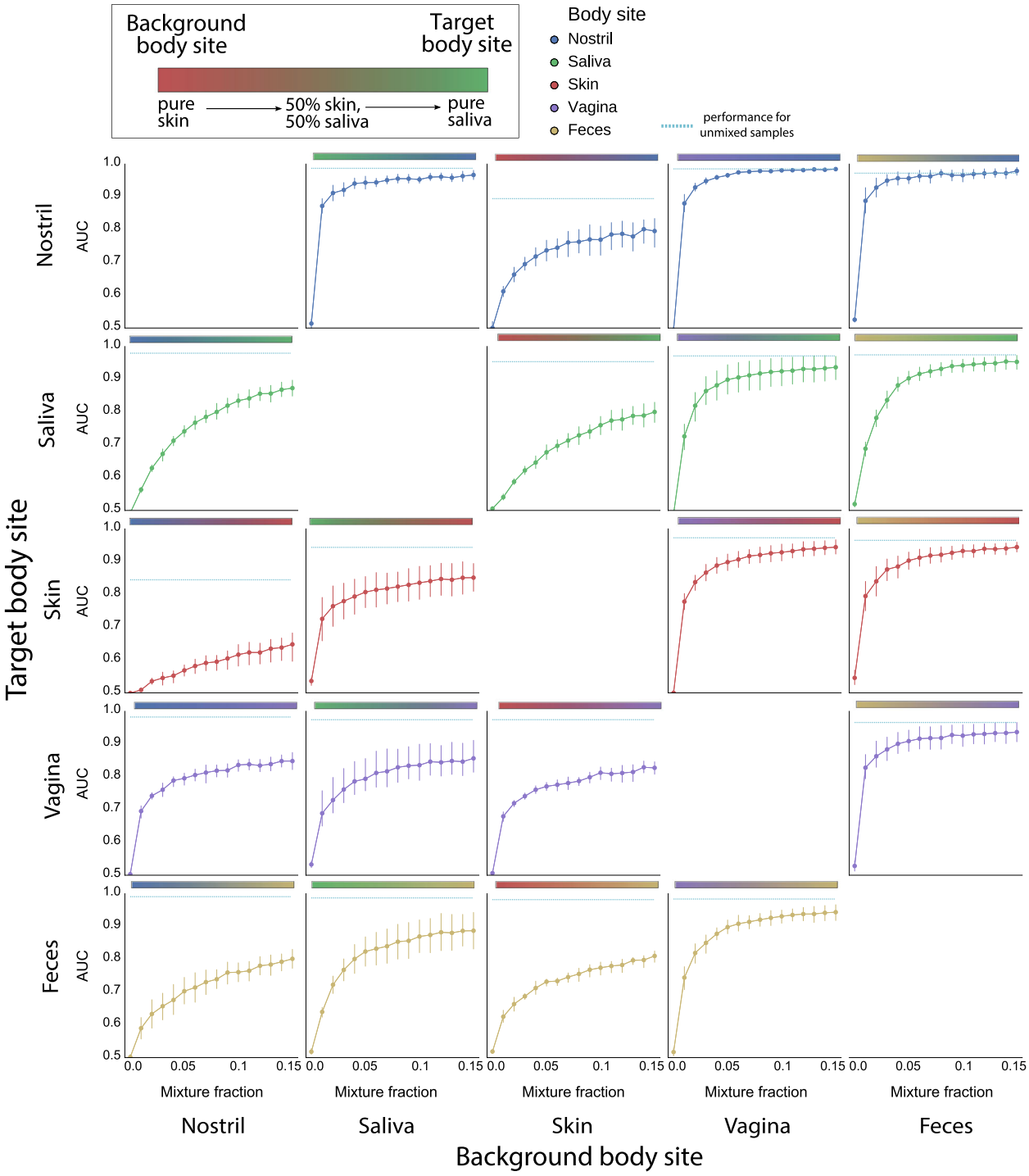


Fig. 3 Discrimination performance of RFC-global on mixed samples with varying levels of mixture fractions (up to 15%). Mixtures of all pairs of body sites were prepared using an in silico procedure (see the "Methods" section) and then predicted by RFC-global. Prediction performance was quantified in terms of AUC. Dotted lines indicate performance evaluated on unmixed samples

biomarker sets more parsimonious, as OTUs found to be only indirectly associated with a habitat are excluded.

In our study, GLL reduced the number of OTUs from an initial 60,982 to 635 directly associated biomarkers, between 92 (nostril) and 326 (skin) (Fig. 4b, Additional file 9: Table S5). When evaluating single-source samples, the predictive performance (F1 score) of a RFC trained only on biomarkers (RFC-global-GLL) was slightly increased (1 to

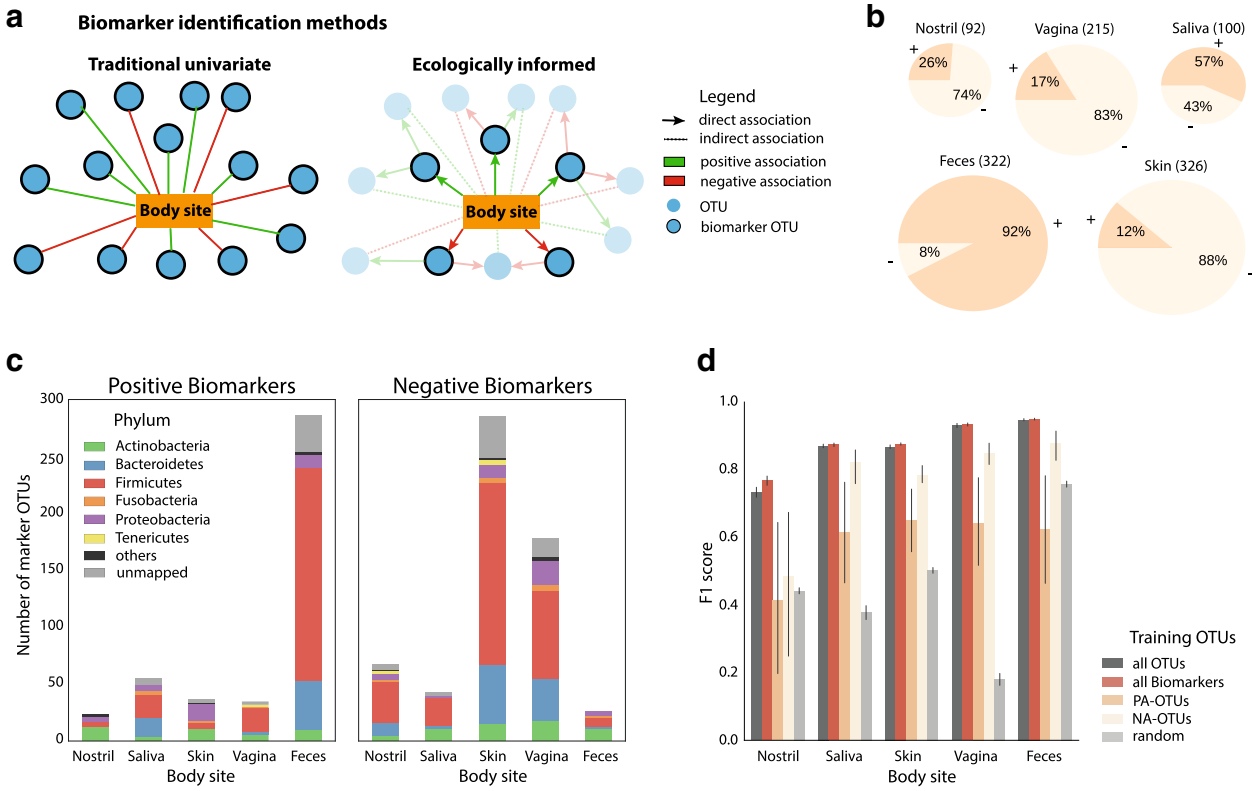


Fig. 4 Identified ecologically informed biomarker OTUs. **a** Comparison of traditional univariate and ecologically informed biomarker identification. Ecological association information can be used to discard biomarkers indirectly associated with body sites. **b** Between 92 and 326 positively and negatively associated biomarkers were identified per body site. **c** Phylum-level distribution of positive and negative biomarkers per body site. **d** Predictive performance of classifiers for each body site, trained on different sets of OTUs: all OTUs, all biomarkers, only PA-OTUs (positive biomarkers), only NA-OTUs (negative biomarkers), and a random set of OTUs of size equal to the all biomarkers

5%) compared to RFC-global (Fig. 4d). In contrast, a random subset of all OTUs with size equal to the biomarker set resulted in strongly reduced F1 scores that ranged from a 20% reduction in feces up to an 80% reduction in vaginal sites (Fig. 4d).

We further evaluated the performance differences between RFC-global versus RFC-global-GLL in simulated mixtures of body site and soil samples and found that biomarker-only RFC models performed worse on such mixtures, with decreases that ranged between 50% AUC for feces and 5% for vaginal sites (Additional file 8: Figure S5).

In order to test how GLL biomarker discovery was affected by dataset size, we further applied it to the Costello et al. dataset. This test yielded 12 biomarkers (98% less than GlobalBodysites) which achieved an average F1 score of 0.39 (compared to 0.89 in RFC-global-GLL, Additional file 10: Figure S6a).

Negatively associated microbes are numerous and contribute strongly to sample prediction accuracy

Among the directly associated biomarkers selected by GLL, we found both positively associated and negatively associated OTUs (PA-OTUs and NA-OTUs, Fig. 4a, b). The presence of PA-OTUs associated with a body site in a sample increased the likelihood of classification as that body site, while their absence decreased it (vice versa for NA-OTUs). We observed a large fraction of PA-OTUs for feces (92%), while for nostril, skin, and vagina samples, mostly NA-OTUs were identified (74 to 88%). For saliva, fractions of positively and negatively associated OTUs were balanced (57% PA-OTUs). Average prevalence—the number of samples a biomarker OTU is found in across all body sites—was slightly elevated for NA-OTUs (Additional file 11: Figure S9).

All pairs of body sites showed a varying amount of overlap among identified biomarker OTUs (Additional file 12: Figure S7). Notably, this overlap was in most cases oppositional (positive for one body site, negative for the other). An exception to this pattern was the combination nostril-skin, for which 98% of shared biomarker OTUs had the same association type.

We compared the importance of the PA-OTUs and NA-OTUs in classification accuracy by training RFC classifiers separately on only positively or negatively associated biomarkers. When using PA-OTUs only, we

observed 25–55% lower F1 scores (for skin and nostril, respectively). In contrast, using only NA-OTUs resulted in smaller decreases in F1 score ranging between 4% for saliva to 11% for skin for all sites except nostril which showed a reduction of 35% (Fig. 4d).

Previously unreported associations between microbes and body sites

We next examined the taxonomic profiles of detected biomarker OTUs. These OTUs stemmed almost exclusively from the domain *Bacteria*, with a small number of *Archaea* and *Eukaryota* (0.3% and 0.5% respectively). As shown in Fig. 4c, four phyla were dominant in the selected biomarkers across all body sites: *Bacteroidetes*, *Firmicutes*, *Proteobacteria*, and *Actinobacteria*. With few exceptions, both PA-OTUs and NA-OTUs from these phyla were found across all body sites. We found high fractions of positive *Firmicutes* and *Bacteroidetes* biomarkers in feces and *Proteobacteria* biomarkers in skin. Furthermore, *Tenericutes* included predominantly PA-OTUs for vagina and NA-OTUs for skin and nostril.

We identified 107 distinct genera among all 635 biomarker OTUs. Numerous associations found in our automated analysis were in line with previously reported associations (e.g., *Corynebacterium* and *Cutibacterium* for skin, *Lactobacillus* for vagina, *Bacteroides* for feces, and *Prevotella* for saliva; see Additional file 13: Table S1 and Additional file 14: Table S2). Nonetheless, we also detected novel specific genus-body site associations that, to our knowledge, have not been discussed elsewhere (Table 1): *Ralstonia* and *Caulobacter* were found to be directly associated with skin, *Delftia* to nostril, and an archaeal OTU mapping to the genus *Halovenus* to feces. It is furthermore noteworthy that GLL marked a number of previously reported genera as only indirectly associated in our study, for instance *Megasphaera* in saliva [32], *Veillonella* in skin [26], *Mobiluncus* in vagina [33], and *Bacillus* in feces [34].

Across all genera that comprised the 635 identified biomarker OTUs, 14 included PA-OTUs for multiple body sites. For instance, *Actinomyces* contained PA-OTUs for saliva, skin, and vagina. Moreover, five genera included both PA-OTUs and NA-OTUs for the same body site (mostly vagina, Additional file 15: Table S3). For example, *Prevotella* contained three PA-OTUs and six NA-OTUs for vagina.

Table 1 Novel positively associated biomarker genera for each body site. Weight of each biomarker is measured by the percentile of feature importance in RFC-global amongst biomarker OTUs of the same body site. Prevalence is the average number of samples that biomarker OTUs of a genus were found in across all body sites

Body site	Genus	Prevalence	Percentile of Random Forest feature importance
Nostril	*Delftia*	2480	83.0
Skin	*Ralstonia*	2666	83.7
	Caulobacter	1280	71.2
Feces	Putative *Halovenus*	266	0.1

To gain an understanding of the taxonomic relationships among selected biomarkers, we reconstructed a phylogenetic tree for a set of 50 OTUs categorized as the most informative for classification by the Random Forest models (Fig. 5). While most clades at varying tree depths displayed a strong positive association to a single body site, many also included members with shifted habitat preference. For instance, nearly all OTUs in the *Bacteriodetes* clade (phylum level) were positively associated exclusively with saliva. However, one *Prevotella* OTU from this clade was instead positively associated with vaginal sites, and one *Bacteroides* OTU with feces. Similarly, one OTU in the *Atopobium* clade (genus level) was a positive biomarker for saliva while the second one was positively associated with vaginal sites. Overall, association patterns were particularly pronounced for the *Actinobacteria* clade (phylum level): 29% of its members were directly positively or negatively associated to four or more body sites (average associations per OTU, 2.8). In contrast, the *Firmicutes* clade had only 10% OTUs with four or more direct associations (average, 2.1).

Critically, 58% of all biomarker OTUs could not be mapped to any known genus and many of these microbes were among the most important biomarkers (Fig. 5, Additional file 16: Figure S8). Out of this set, 21% could only be confidently classified at the domain or phylum level, with two members surpassing the 90th percentile of feature importance (mean, 40th) (Additional file 16: Figure S8). Furthermore, many of these largely unclassified biomarkers were common: on average, they were found in 1012 samples (up to 4542). In order to characterize the taxonomic neighborhood of biomarkers not confidently mapping to any phylum, we further analyzed their closest 16S rRNA matches and found that most of these OTUs hit *Firmicutes* (55%), *Proteobacteria* (17%), *Bacteroidetes* (11%), and *Tenericutes* (10%), albeit at low sequence identity. Among these phyla, we observed an overrepresentation of unclassified OTUs for *Proteobacteria* (17% unclassified compared to 8% classified) and *Tenericutes* (10% compared to 0.6%).

Aerobicity is the most defining characteristic of microbial biomarkers found in body sites

To further characterize the selected microbial biomarkers, we collected information on oxygen dependency, shape, gram stain, spore formation, and motility in a literature search and tested which of these microbial traits were enriched among the selected biomarkers at

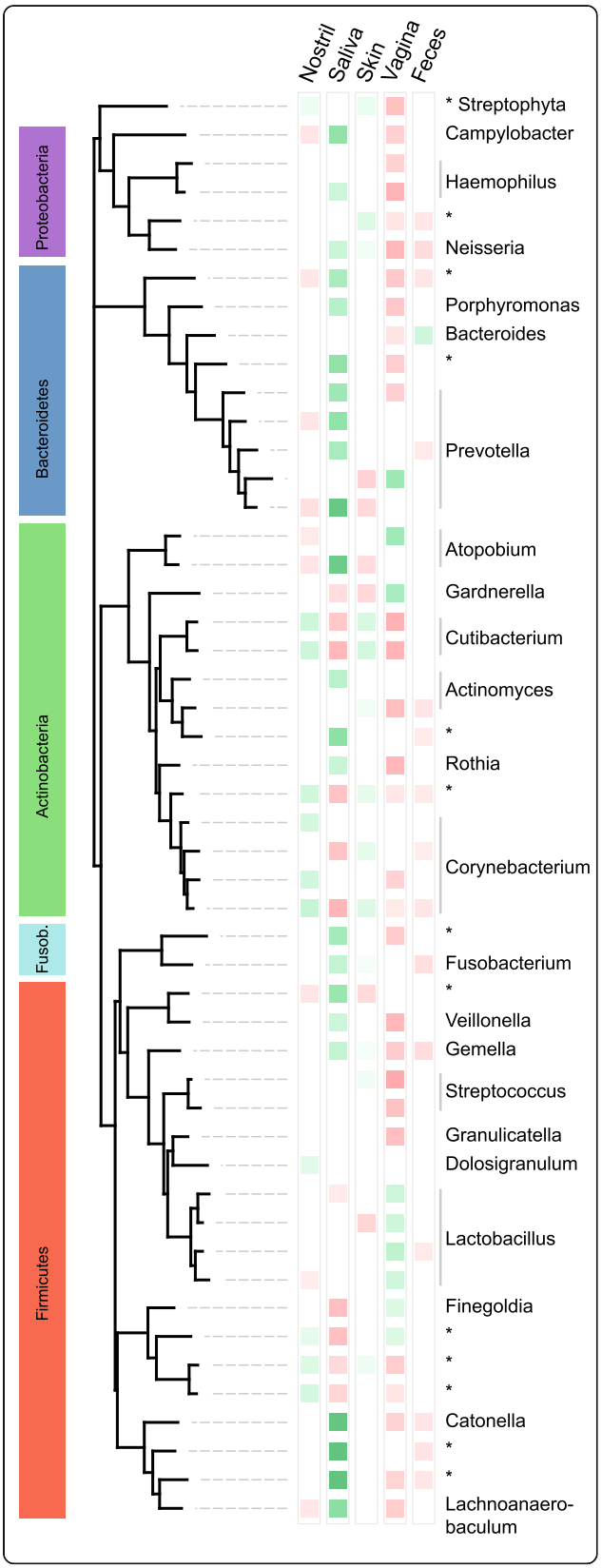


Fig. 5 Phylogenetic distribution of the top 50 biomarkers. Biomarker OTUs were chosen based on feature importance (gini impurity) as estimated by RFC-global. Colored blocks represent biomarker association type and strength, measured as normalized mutual information, where green indicates positive and red negative association. OTUs unclassified at the genus level are denoted by "*". The shortened phylum name stands for *Fusobacteria*

each body site compared to the background of all other biomarkers with the same association type (Table 2). We found the most specific enrichment pattern for PA-OTUs in feces: these microbes were significantly enriched for anaerobes and tended to be rod shaped, spore forming, and motile. In contrast, NA-OTUs for feces tended to be facultative anaerobes or gram positive.

Across body sites, we found a clear separation by oxygen dependency, where oxygenated body sites (nostril, skin) were enriched for aerobic or facultative anaerobic biomarkers, while feces were enriched for anaerobes. Vagina, being a less oxygenated environment, showed an enrichment of facultative anaerobes only, while the well-oxygenated saliva environment had no significant signal for oxygen dependency.

Discussion

This study is, to our knowledge, the most comprehensive cross-study evaluation of human body site classification and the first analysis of ecologically informed biomarkers for human body sites. Our results show that aggregating a large number of microbial sequences from diverse studies on human body sites leads to (i) strongly improved classification of body sites and (ii) the identification of parsimonious and likely less biased sets of microbial biomarkers for body sites. In our evaluation of body site classification, we highlight the prediction performance achieved for the detection of mixture components. This characteristic of our classification model is particularly valuable for accuracy-demanding applications as for instance in forensics. Limitations of our study include the observation that the classification of very similar sample types, such as nostril and skin, remains challenging. Moreover, the directly associated microbial biomarkers we report here require further experimental validation.

Improved classification accuracy in large cross-study datasets

We analyzed a large-scale dataset composed of over 15,000 samples from five human body sites and showed that a RFC model trained on this data (RFC-global) is considerably more accurate for body site prediction than multiple models trained on data from single studies (RFC-single, RFC-single-hmp) used in previous body site

Table 2 Enriched physiological traits among biomarker OTUs by body site and association type

Body site	Association type	Physiological trait
Nostril	+	Aerobic, facultative anaerobic, gram positive
	–	No enriched traits
Saliva	+	Spherical shape
	–	Gram positive
Skin	+	Aerobic, facultative anaerobic
	–	Anaerobic
Vagina	+	Facultative anaerobic
	–	No enriched traits
Feces	+	Anaerobic, rod shaped, spore forming, motile
	–	No enriched traits

classification benchmarks [10, 11, 25]. We found that prediction performance continued improving when including additional studies (after correction for study size), indicating that even more data from undersampled geographic regions or conditions could be useful to further improve the performance of RFC-global. Moreover, we demonstrated that in mixtures comprising microbial communities from two human body sites or a human body site and a soil sample, RFC-global is capable of detecting trace amounts (down to 1% mixture fraction) of a target sample type in many tested cases—this sensitivity was not achieved with RFC-single.

The discrimination power demonstrated here is useful for a number of applications. In a forensic case involving sexual assault, for example discerning whether a stain at a crime scene contains a mixture of vaginal fluid and skin (from different individuals) or whether it contains vaginal fluid and saliva can affect the reconstruction of the crime event. Similarly, the power to discriminate human body site components from soil could help trace body site samples in forensic stains that have been exposed to environmental bacteria for prolonged periods of time, for example in forest environments. Of particular interest is that the high classification performance in mixtures with soil samples was achieved without prior training on soil communities. This inherent robustness of RFC-global to soil-based noise may thus potentially extend to other environments. Since many potential source environments for microbial transfer may still be unknown or undersampled, generic robustness to noise sources would be an important feature. Albeit further confirmation is needed, this also indicates promise for similar classification tasks, like identification of human sewage pollution in water and indoor contamination (or microbial transfer) on hospital surfaces.

As a cautionary note, we observed that classification accuracy and the detection limit in mixtures depended strongly on sample type. For example, sample types harboring relatively similar microbial communities, such as nostril and skin or, to a lesser extent, feces and soil, were harder to distinguish and resulted in more misclassifications and higher detection limits than more distinct sample types, such as nostril and feces. Only high-confidence mixture classifications for similar sample types should thereby be trusted.

A core set of ecologically informed biomarkers

Applying Generalized Local Learning (GLL) enabled removal of redundant biomarkers whose association was statistically explainable by microbe-microbe associations. This selection step led to a strongly reduced core set of microbial markers that are directly associated with human body sites. We showed that this set precisely captured body site differences, achieving classification accuracy similar to or surpassing the full set of OTUs in unmixed samples. Applying GLL on the Costello et al. dataset [13] identified a reduced set of 12 biomarker OTUs (compared to 635 in the global dataset), resulting in a sharp drop in classification performance (Additional file 10: Figure S6). This drop likely stems from the smaller size and reduced diversity of the Costello et al. dataset, which emphasizes the need for heterogeneous large-scale datasets to fully take advantage of GLL.

We examined two types of biomarkers for each body site: positively associated (PA-OTUs) and negatively associated (NA-OTUs) (Fig. 4b). In contrast to most other body sites, feces were characterized by a larger number of PA-OTUs and few NA-OTUs. This trend was likely a consequence of the distinctness of gut communities from other body sites, as most of these positive, feces-specific biomarkers were identified as negative for at least one other body site (Additional file 12: Figure S7).

Although NA-OTUs are commonly reported by LEfSe [35]—a standard tool for microbial biomarker discovery that does not distinguish between direct and indirect interactions—negative association patterns between microbial taxa and human body sites have to our knowledge not been comprehensively discussed in previous literature. We showed that NA-OTUs are numerous, can achieve levels of accuracy comparable to the use of both OTU types, and result in consistently higher predictive performance than using PA-OTUs alone (Fig. 4d). While NA-OTUs are generally more numerous than PA-OTUs, in particular in nostril, skin, and vagina, this cannot explain observed performance differences, as NA-OTUs also outperform PA-OTUs in body sites with lower NA-OTU proportions. The predictive superiority of NA-OTUs is striking because it indicates that the absence of specific microbial taxa is generally more

informative than the presence of usual microbial taxa in an environment. We expect ecological factors driving this strong negative association signal to be mostly non-microbial (e.g., pH, oxygen content, medication), since our GLL analysis reduced the influence of inhibiting ecological microbial associations, such as competition and amensalism. Similarly, positive associations have been corrected for symbiosis and commensalism, leaving unmeasured non-microbial variables as the most likely explanation for observed positive biomarkers. Identifying factors driving the direct associations we report and pinpointing them to particular microbes and body sites would provide important insights into the forces shaping microbial diversity across the human body.

Taxonomic and phylogenetic patterns of detected biomarkers

The importance of using large aggregated datasets and selecting ecologically informed biomarkers is highlighted in the taxonomic analyses of identified markers. While many biomarker OTUs identified here belonged to previously reported site-specific genera, we also found novel associations. Furthermore, we found that some associations previously reported in single studies were not confirmed in our study. It is worth noting that the conditioning step of GLL can only exclude (but never add) biomarkers; thus, any novel biomarkers found in this study result from the increased size and diversity of the analyzed dataset. Notable examples of novel markers are *Ralstonia* and *Caulobacter* OTUs for skin (Table 1) and an OTU mapping to the archaeal genus *Halovenus* as a biomarker for feces. The latter is surprising because mostly water-dwelling, halophile species of this genus have been described thus far. Since archaeal diversity is underrepresented in current taxonomic databases, it is thus possible that this weakly predictive and low-prevalence OTU belongs to a to-date unidentified archaeal genus able to persist in the human gut.

A number of commonly reported genus associations were not found as direct associations in our analysis. For example, while *Methanobrevibacter* [36] is one of the few archaeal genera consistently identified in the human gut, its known ecological dependence on fermenting bacteria [37] lead to its exclusion as direct biomarker in our study. Similarly, *Veillonella* has been excluded as a biomarker for skin because of its statistical association with a *Streptococcus* OTU, and indeed, symbiosis between *Veillonella* and *Streptococcus* has been previously described [38].

We observed that most major human-associated phyla included both positive and negative biomarkers for all body sites, making high-level taxonomic affiliation only weakly indicative of body site presence. Even at the genus level, we find frequent cases of genera that include biomarker OTUs of the same association type for different body sites. For example, while most *Prevotella* markers were positively associated with saliva, one sub-clade in the phylogenetic tree was associated with vagina (Fig. 5). Furthermore, some genera included PA-OTUs for as many as three distinct body sites, and 5% of all biomarker genera include both PA- and NA-OTUs for the same body site, making these genera highly unreliable for body site classification. We therefore generally recommend analysis at 96% OTU-level resolution or higher to identify predictive biomarkers. A caveat to this approach is that association patterns specific to more general taxonomic levels can be missed. For instance, *Staphylococcus* as a biomarker for skin was not recovered in our analysis because many reads mapping to this genus could not be confidently assigned to one single 96% OTU (see Additional file 17: Text S1). A hybrid approach of unsupervised OTU clusters and supervised taxon assignments may alleviate this problem in future studies. Furthermore, a number of body site-specific strains have recently been described [39], indicating that a strain-level analysis may lead to additional biomarkers and increased classification accuracy in future studies.

Additionally, we note that many microbial biomarkers, some of which are among the most predictive OTUs, could not be precisely taxonomically classified, constituting "microbial dark matter" [40, 41]. For instance, we identified a bacterial OTU distantly related to *Firmicutes* (81% 16S rRNA sequence identity) among the 10% most important biomarkers. It is a strong PA-OTU for saliva and a NA-OTU for nostril, skin, and vagina. Moreover, we observed an overrepresentation of uncharacterized OTUs distantly related to *Proteobacteria* and *Tenericutes*, indicating insufficient coverage of the phylogenetic tree around these phyla in current taxonomic reference databases. We deem it crucial to intensify research on describing uncharacterized human-associated microbes detected in this study in order to elucidate their potential roles in human health and disease.

Microbial trait enrichment in particular body sites

In terms of physiological traits identified for the microbial biomarkers, we find oxygen dependency to be the most pronounced physiological characteristic among PA and NA biomarkers: aerobic microbes tended to be positively associated with exposed body sites, while (facultative) anaerobic microbes preferred lowly oxygenated sites. Apart from this expected observation, only PA feces biomarkers showed a detailed enrichment pattern for multiple other traits, namely rod shape, motility, and spore formation. Compared to coccoid cells, rod-shaped cells have a higher

surface to volume ratio and are therefore more efficient at uptaking substrate [42, 43]. Along the same lines, motility has been found to be an important feature in competitive environments [44, 45]. The enrichment in microbial species with these traits thus indicates that despite being a nutrient-rich environment, the gut selects for microbial traits providing a benefit in competitive habitats. Furthermore, spore-forming lactic acid bacterial species of the *Lactobacillus* genus have been shown to be sensitive or weakly tolerant when exposed to acidic environments and bile acid [46]. However, in endospore form, these bacteria survive the passage through the human stomach and germinate successfully in the gut [47]. The ability to survive the acidic stomach environment that presents a barrier to the gut would therefore confer a clear advantage to microbial species frequently exposed to such environments and suggests that many of these species could be acquired with food, rather than reside permanently in the gut.

Limitations

Limitations of our study include the use of in silico simulations for mixture analysis. While our protocol uses real samples to create mixtures, it remains to be investigated how mixtures created in a wet lab environment would affect classifier performance. Additionally, we only estimated robustness to environmental noise through soil samples as a typical environmental source. Whether the robustness we observe generalizes to other likely noise sources, such as dust from indoor environments, requires further testing. Importantly, a classifier trained only on GLL-selected biomarkers was markedly less robust to environmental noise, likely because no markers were selected to discriminate this signal. It is therefore crucial to use the classifier trained on all OTUs if environmental mixtures are expected. Finally, the direct microbe-environment associations we report here are based on statistical relationships and whether these reflect real direct environmental dependencies requires experimental validation.

Conclusions

The present study is part of a recent development, where the ongoing growth and diversification of microbial sequencing studies of human body sites, coupled with adequate statistical techniques to mine emerging patterns from this data, catalyzes discovery. We are confident that this data-intense approach will continue to expand our understanding of the human microbiome and lead to generalized insights into our microbial ecosphere.

Methods

Data acquisition

We first selected a set of representative body sites: saliva, skin, vagina, and feces. We further included the body site nostril due to its known similarity with skin, in order to also test classification accuracy on the more difficult distinction between skin and nostril in later analysis steps.

Studies from the NCBI Sequence Read Archive database (SRA, [29]) were filtered for human samples through automated parsing of annotation keywords, matching at least one of the following rules: (1) "Human" or "*Homo sapiens*" is found in the host name field, (2) "9606" is found in either the host taxon ID or sample taxon ID field, or (3) the pattern "human <*> metagenome" is found in the organism field, where "<*>" is a wildcard for a single word (e.g., "gut") or empty. On these filtered human samples, we then conducted a second body site-targeted search with the keywords "saliva," "tongue," "nostril," "nares," "vagina," "fornix," "retroauricular crease," "antecubital fossa," "skin," and "feces." Random subsets of samples for each body site were manually checked via the SRA web service to verify that the filtered samples were of human origin and belonged to the habitats assigned by the automated pipeline. In case of mismatches, samples were removed from the pool. Similarly, soil samples were retrieved from the SRA through the keyword "soil" and subsequently filtered for samples with no body site-related keyword annotations.

OTU mapping, taxonomic classification, and filtering

Raw sequence data for 50,273 sequenced samples (independent sequencing runs of biological samples) from the keyword-filtered studies was downloaded from the NCBI SRA database. Reads were quality filtered using custom programs that trimmed reads to the first two consecutive low-quality bases (≤ 10) and discarded reads smaller than 75 bp or having a fraction of low-quality reads larger than 5%. MAPseq v1.0 [30] was used to map the filtered reads to the reference of full-length 16S/18S rRNA sequences provided with MAPseq which includes representatives for 61,899 OTUs at the 96% identity cutoff. The results of MAPseq were parsed and an OTU count table was created using the assignments to OTUs at 96% sequence identity with a minimum confidence of 0.5. Taxonomy was assigned to OTUs based on a 90% consensus over the full taxonomic lineages of all OTU member sequences. For sequences belonging to RefSeq [48] genomes or culture collection strains, the annotated taxonomy as provided by NCBI in December 2017 was used. Other sequences were taxonomically classified through mapping onto the RefSeq set using MAPseq and a confidence threshold of 0.5.

An initial analysis (see Additional file 17: Text S1) showed that samples with fewer than 20 unique OTUs tended to be noisy; we therefore excluded these from the analysis. We then computed the normalized mutual information between OTUs and performed hierarchical clustering (complete linkage, Euclidean distance) to group highly similar OTUs with normalized mutual information higher than 0.9 together, as these OTUs were hard to distinguish from each other by GLL, the biomarker discovery algorithm we applied. For each group, a random representative was chosen. After all filtering steps, a subset of 15,082 samples remained for analysis.

Classification of body sites

Counts in the OTU table were normalized by the total number of mapped reads per sample, resulting in relative abundances. Next, the dataset was split into five distinct training and validation sets, retaining the proportions of samples across body sites constant for each subset (stratified k-fold split). For each subset, a Random Forest Classifier (RFC, [15]) was trained with the python package scikit-learn [49]. Hyper parameters were optimized based on a grid search across a stratified inner fourfold cross-validation on each respective training set. Possible parameter values were as follows: number of trees 500, 1000, or 2000; maximum number of features $\frac{1}{2} \cdot \sqrt{\text{features}}$, $\sqrt{\text{features}}$, $2 \times \sqrt{\text{features}}$. Class weights were automatically adjusted to cope with class imbalance, and the optimization objective for the inner cross-validation loop was the F1 score. Classifier output probabilities were additionally calibrated through a non-parametric procedure based on isotonic regression [50] implemented in scikit-learn. All five classifiers were evaluated on their respective validation sets to estimate the generalization error of the final classifier trained on the whole dataset (RFC-global).

For RFC-single and RFC-single-hmp, the same procedure was slightly adapted to use only samples from the study by Costello et al. [13] or the Human Microbiome Project [26] for training. Furthermore, vaginal samples were excluded from the validation sets for the comparison between RFC-global and RFC-single since samples of this type were not collected in [13].

For "weighted" F1 score analyses, we weighted each sample by the inverse of the number of samples belonging to its associated study. These weights were then passed as "sample_weight" parameter in the scikit-learn F1 score function.

Biased validation and training sets were created by keeping all samples from one body site (B_{bias}) and then randomly down-sampling all other body sites until each had 10% of the sample count of B_{bias}.

Mixture simulations

Artificial sample mixtures from two body sites B_{target} and $B_{\text{background}}$ were created through the following procedure: (i) randomly choose one sample taken from B_{target} and $B_{\text{background}}$ respectively; (ii) compute relative frequencies of each OTU in these samples and use the frequencies as base probabilities of OTU-drawing; (iii) weight the probabilities according to the desired mixture fraction $F_{\text{background}}$, which determines how similar the mixed sample should be to the sample from $B_{\text{background}}$; (iv) randomly choose OTU sequences based on these weighted probabilities until n sequences were chosen, where n is the weighted average of sequences in the two samples, with weights $F_{\text{background}}$ and $1 - F_{\text{background}}$. The procedure was repeated for each pair of body sites along a gradient of increasing mixture fractions.

To estimate performance, the data was split into the same training and validation sets as described previously for the classification of unmixed samples. For each of these splits, the classifiers previously trained on the respective training sets were used, while validation sets were further processed, separately for each body site pair B_{target} and $B_{\text{background}}$. For this processing, samples in the validation sets were first reduced to only samples from B_{target} and $B_{\text{background}}$, followed by in silico mixture of all B_{target} validation samples with randomly picked $B_{\text{background}}$ validation samples, using increasing mixture fractions $F_{\text{background}}$. AUC scores were finally computed for each mixed validation set and pre-trained classifier.

We also determined thresholds for the correct identification of a target body site in a mixture. To do so, we first prepared in silico mixtures as described above for all training samples. Then we computed a precision-recall curve for these mixed training samples and picked the threshold that yielded the optimal F1 score on that curve. Subsequently, F1 scores were used to quantify the threshold-adjusted prediction performance of our Random Forest models on mixed test sets (see previous paragraph). This procedure was repeated for each combination of body sites and mixture fractions, leading to a threshold table with $5 \times 4 \times 10 = 200$ entries. For this analysis, only thresholds were adjusted; the classifier decision trees and calibrator were not trained on mixed samples.

Identification of microbial biomarkers

While Random Forest Classifiers perform intrinsic feature selection which can yield insights into which OTUs the classifier estimates to be most predictive [11, 12, 15], this approach has a number of shortcomings. Firstly, if features are highly correlated, the classifier tends to arbitrarily pick one of them and discard the others, leading to the removal of potentially biologically interesting OTUs. Since microbes interact with each other and live

in complex ecological networks of mutual dependencies, such correlations are inevitable. Furthermore, deciding on a cutoff for how many OTUs to label as biomarkers based on feature importance (gini impurity in our case) can be difficult.

To identify the core set of microbial markers directly associated with a body site, we applied Generalized Local Learning (GLL, [27]) to address these shortcomings. The approach detects OTUs whose association with a habitat cannot be statistically explained by their relationship with other microbes, effectively exploiting ecological dependencies among OTUs to make biomarker discovery more parsimonious. Furthermore, it internally uses statistical tests of independence, which apply well-studied significance cutoffs, and avoids classifier-specific inductive biases [31].

GLL was instantiated with semi-interleavedHITON-PC as edge-finding algorithm and mutual information as test metric (proportional to the classic *G*-test, [51]). We ran the algorithm with the following parameters: max-k = 3, h-ps = 5, and alpha = 0.05. Prior to biomarker discovery, OTU abundances in the OTU table were binarized, where an OTU was assigned the value 1 if at least one sample read mapped to the OTU and 0 if not, in order to allow discretized mutual information tests and reduce sequencing depth biases. We ran GLL separately for each body site, using a custom implementation in the python [52] and cython [53] programming languages. False discovery rate adjustment of p values [54] was applied prior to the GLL conditioning step. Whether an OTU was positively or negatively associated with a body site was estimated by the sign of the Spearman correlation coefficient between binarized OTU and body site. We found identical association type assignments for odds ratios and linear discriminant analysis effect sizes.

Phylogenetic analysis of biomarkers

The biomarkers identified by GLL were weighted based on feature importance (gini impurity) inferred by RFC-global, and the top 50 most important biomarkers as estimated by feature importance (gini impurity) were chosen for phylogenetic analysis. A multiple sequence alignment for the selected biomarkers was extracted as a subset of the publicly available alignment of all OTUs in the reference database, created with INFERNAL version 1.1.2 [55] and microbial secondary structure model SSU-ALIGN [56]. Based on this alignment, a phylogenetic tree was reconstructed using fasttree version 2.1.3 [57] using the GTR substitution model and otherwise default options.

Collection of microbial trait information

Across all PA and NA biomarker OTUs, we created a list of genera and reviewed primary literature and the public database MicrobeWiki [58] to assign a list of phenotypic characteristics to them. Major categories were aerobicity (subcategories: aerobe, anaerobe, facultative anaerobe), gram stain (gram positive, gram negative), cell shape (rod, spherical, helical), spore formation (forms spores, does not form spores), and motility (motile, non-motile). When no information was found, the trait was labeled as missing, while if more than one sub category was described by different sources, all alternatives were kept and used later for the statistical analysis. This trait information was then extrapolated to all OTUs with mapped genus information.

Statistical analysis of microbial traits

For each marker OTU subset S from each combination of body site and association type, as well as each trait sub-category (e.g., anaerobic), we tested whether the sub-category was significantly enriched within S compared to the background of marker OTUs with the same association as S, but associated to a different body site. To this end, we conducted Fisher's exact test (alpha 0.05, one-tailed), followed by false discovery rate adjustment of p values [54].

Additional files

Additional file 1: Table S4. Accessions for sequencing runs, sequencing samples, and projects contained in GlobalBodysites. (TSV 440 kb)

Additional file 2: Figure S10. Additional performance comparisons of RFC-global. (A) Comparison of RFC-global to a classifier trained on the Human Microbiome Project subset of GlobalBodysites (RFC-single-hmp). "unweighted": default F1 scores are computed, without applying weights; "weighted": samples are weighted inversely to the size of the study they belong to for F1 score calculation, resulting in a penalty for large studies and higher importance of smaller studies. (B) Robustness of RFC-global and two of its variations to body site proportion biases in the validation sets. Test set definitions are as follows: "*-biased", body site "*" had 10× more samples in the validation set than other body sites; "Equal", all body sites had equal proportions (equivalent to the site with the fewest samples). Classifiers are: "RFC-global", the original RFC-global classifier; "RFC-global-balanced", trained on equal body site proportions; "RFC-global-Feces-biased", trained on a biased set with 10× more feces samples than other body sites. (PDF 426 kb)

Additional file 3: Figure S1. Comparison of OTU importance between RFC-global and RFC-single. (A) Overlap of predictive OTUs (feature importance > 0) between classifiers, (B) joint distribution of unpredictive (U.P.) and predictive (P.) OTUs, and (C) direct comparison of feature importances for 1277 OTUs predictive in both classifiers. (PDF 626 kb)

Additional file 4: Figure S11. Improvement of RFC-global performance with increasing numbers of studies and samples. "unweighted": default F1 scores are computed without applying weights; "weighted": samples are weighted inversely to the size of the study they belong to for F1 score calculation, resulting in a penalty for large studies and higher importance of smaller studies. (A) Performance in relation to increasing numbers of studies, starting with only the HMP dataset. (B) Re-mapping of (A) to the numbers of samples included in each set of studies. (PDF 413 kb)

Additional file 5: Figure S2. Discrimination performance of RFC-global on mixed samples. Along a gradient of increasing mixture fractions (0 to 100%), unseen samples for all pairs of body sites were combined into mixed samples using an in silico procedure (see the "Methods" section) and then predicted by RFC-global. Prediction performance was quantified in terms of AUC. (PDF 508 kb)

Additional file 6: Figure S4. Comparison of discrimination performance between RFC-global and RFC-single on mixed samples. Along a gradient of increasing mixture fractions (0 to 100%), unseen samples for all pairs of body sites were combined into mixed samples using an in silico procedure (see the "Methods" section) and then predicted by both classifiers. Prediction performance was quantified in terms of AUC. (PDF 456 kb)

Additional file 7: Figure S3. Discrimination performance of RFC-global on mixed samples. Along a gradient of increasing mixture fractions (0 to 100%), unseen samples for all pairs of body sites were combined into mixed samples using an in silico procedure (see the "Methods" section) and then predicted by RFC-global (thresholds optimized on the training sets). Prediction performance was quantified in terms of F1 score. (PDF 501 kb)

Additional file 8: Figure S5. Comparison of discrimination performance between RFC-global with all OTUs and only biomarker OTUs on samples contaminated with soil. Along a gradient of increasing mixture fractions (0 to 100%), unseen samples of each body site were contaminated with soil using an in silico procedure (see the "Methods" section) and then predicted by RFC-global, using each respective OTU set. Prediction performance was quantified in terms of AUC. (PDF 375 kb)

Additional file 9: Table S5. Comprehensive information of all biomarker OTUs, including marker association types, marker strength, and meta-information such as NCBI 16S accessions and taxonomic lineages. (TSV 68 kb)

Additional file 10: Figure S6. Prediction performance across body sites for RFC-global vs. RFC-single, as well as all OTUs vs. only biomarker OTUs. Ecologically informed biomarker OTUs were extracted from the whole GlobalBodysites dataset for RFC-global and from a single-study subset [13] for RFC-single. Prediction performance was measured as (A) AUC and (B) F1 score. (PDF 389 kb)

Additional file 11: Figure S9. Comparison of prevalence and mean relative abundance between NA-OTUs, PA-OTUs (strictly positive, PA; strictly negative, NA), and non-biomarkers. Mean relative abundances were scaled by taking their square root. (A) Sample quantities calculated across all samples. (B) Sample quantities stratified by body site, where PA- and NA-OTUs only include biomarkers for each respective body site. (PDF 3389 kb)

Additional file 12: Figure S7. Overlaps of identified biomarker OTUs between body sites. Diagonal: distribution of PA-OTUs and NA-OTUs for each body site. Upper triangle: pairwise overlaps of biomarker OTUs for each body site, quantified by Jaccard similarity (indicated by "Jacc"). Lower triangle: normalized joint distribution of PA-OTUs and NA-OTUs for each body site pair. High values indicate large fractions of OTUs with one association type in the first row body site (rows) and a second association type in the second body site (columns). (PDF 816 kb)

Additional file 13: Table S1. Top 5 most important positively associated biomarker genera per body site, measured by maximum feature importance in RFC-global among biomarker OTUs of each genus. (DOCX 7 kb)

Additional file 14: Table S2. Novel and previously described positive associations between genera and body sites. (DOCX 23 kb)

Additional file 15: Table S3. Numbers of NA- and PA-OTUs per genus and body site. (TSV 5 kb)

Additional file 16: Figure S8. Feature importances by taxonomic classification quality. For each taxonomic rank, shows feature importances of all biomarker OTUs confidently classified down to that rank, but not further. (PDF 343 kb)

Additional file 17: Text S1. Supplementary methods and analyses. (DOCX 21 kb)

Abbreviations

AUC: Area under the ROC curve; GLL: Generalized Local Learning; NA-OTU: Negatively associated biomarker OTU; OTU: Operational taxonomic unit; PA-OTU: Positively associated biomarker OTU; RFC: Random Forest Classifier; ROC: Receiver operating characteristic; SRA: Sequence Read Archive; WGS: Whole genome sequencing

Acknowledgements

We thank Akos Dobay for the helpful discussions and feedback on this manuscript.

Funding

This work was supported by the Swiss National Science Foundation (grant no. 31003A-160095).

Availability of data and materials

Short read sequencing data used for GlobalBodysites can be downloaded from the NCBI Sequencing Read Archive database [29] using the accession numbers provided in Additional file 1: Table S4. Pre-trained classification models are provided by the authors upon request.

Authors' contributions

JT, NA, JFMR, and CvM conceived the project. JT, TSBS, and JFMR contributed to data collection and preprocessing. JT contributed to classifier training, biomarker detection, and downstream analysis. JT and JFMR contributed to figures. All authors contributed to data interpretation. JT, NA, and JFMR wrote the manuscript. All authors reviewed the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Author details

[1]Institute of Molecular Life Sciences and Swiss Institute of Bioinformatics, University of Zurich, Zurich, Switzerland. [2]Zurich Institute of Forensic Medicine, University of Zurich, Zurich, Switzerland. [3]Present address: European Molecular Biology Laboratory, Heidelberg, Germany.

Received: 20 April 2018 Accepted: 28 September 2018
Published online: 24 October 2018

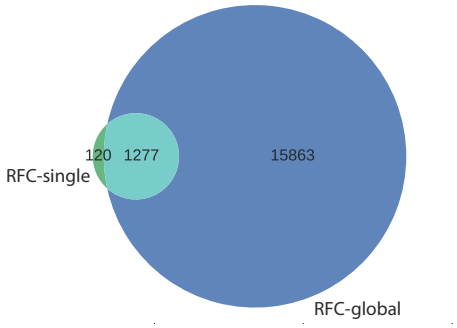
References

1. Zeller G, Tap J, Voigt AY, et al. Potential of fecal microbiota for early-stage detection of colorectal cancer. Mol Syst Biol. 2014;10:766.
2. Mira-Pascual L, Cabrera-Rubio R, Ocon S, et al. Microbial mucosal colonic shifts associated with the development of colorectal cancer reveal the presence of different bacterial and archaeal biomarkers. J Gastroenterol. 2015;50:167–79.
3. Tong M, Li X, Wegener Parfrey L, et al. A modular organization of the human intestinal mucosal microbiota and its association with inflammatory bowel disease. PLoS One. 2013;8:e80702.
4. Yoshizawa JM, Schafer CA, Schafer JJ, et al. Salivary biomarkers: toward future clinical and diagnostic utilities. Clin Microbiol Rev. 2013;26:781–91.
5. Choi A, Shin K-J, Yang WI, et al. Body fluid identification by integrated analysis of DNA methylation and body fluid-specific microbial DNA. Int J Legal Med. 2014;128:33–41.
6. O'Mullan GD, Elias Dueker M, Juhl AR. Challenges to managing microbial fecal pollution in coastal environments: extra-enteric ecology and microbial exchange among water, sediment, and air. Curr Pollut Rep. 2017;3:1–16.
7. Fisher JC, Eren AM, Green HC, et al. Comparison of sewage and animal fecal microbiomes by using oligotyping reveals potential human fecal indicators in multiple taxonomic groups. Appl Environ Microbiol. 2015;81:7023–33.
8. Turnbaugh PJ, Ley RE, Hamady M, et al. The human microbiome project. Nature. 2007;449:804–10.
9. Arumugam M, Raes J, Pelletier E, et al. Enterotypes of the human gut microbiome. Nature. 2011;473:174–80.
10. Knights D, Costello EK, Knight R. Supervised classification of human microbiota. FEMS Microbiol Rev. 2011;35:343–59.

11. Statnikov A, Henaff M, Narendra V, et al. A comprehensive evaluation of multicategory classification methods for microbiomic data. Microbiome. 2013;1:11.
12. Statnikov A, Alekseyenko AV, Li Z, et al. Microbiomic signatures of psoriasis: feasibility and methodology comparison. Sci Rep. 2013;3:2620.
13. Costello EK, Lauber CL, Hamady M, et al. Bacterial community variation in human body habitats across space and time. Science. 2009;326:1694–7.
14. Fierer N, Lauber CL, Zhou N, et al. Forensic identification using skin bacterial communities. Proc Natl Acad Sci U S A. 2010;107:6477–81.
15. Breiman L. Random forests. Mach Learn. 2001;1:5–32.
16. Beck D, Foster JA. Machine learning techniques accurately classify microbial communities by bacterial vaginosis characteristics. PLoS One. 2014;9:e87830.
17. Yatsunenko T, Rey FE, Manary MJ, et al. Human gut microbiome viewed across age and geography. Nature. 2012;486:222–7.
18. Yang C, Mills D, Mathee K, et al. An eco-informatics tool for microbial community studies: supervised classification of amplicon length heterogeneity (ALH) profiles of 16S rRNA. J Microbiol Methods. 2006;65:49–62.
19. Lax S, Hampton-Marcell JT, Gibbons SM, et al. Forensic analysis of the microbiome of phones and shoes. Microbiome. 2015;3:21.
20. Knights D, Kuczynski J, Charlson ES, et al. Bayesian community-wide culture-independent microbial source tracking. Nat Methods. 2011;8:761–3.
21. Tremblay J, Singh K, Fern A, et al. Primer and platform effects on 16S rRNA tag sequencing. Front Microbiol. 2015;6:771.
22. Klindworth A, Pruesse E, Schweer T, et al. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic Acids Res. 2013;41:e1.
23. Gohl DM, Vangay P, Garbe J, et al. Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies. Nat Biotechnol. 2016;34:942–9.
24. Duvallet C, Gibbons SM, Gurry T, et al. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. Nat Commun. 2017;8:1784.
25. Pasolli E, Truong DT, Malik F, et al. Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. PLoS Comput Biol. 2016;12:e1004977.
26. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. Nature. 2012;486:207–14.
27. Aliferis CF, Statnikov A, Tsamardinos I, et al. Local causal and Markov blanket induction for causal discovery and feature selection for classification part i: algorithms and empirical evaluation. J Mach Learn Res. 2010;11:171–234.
28. Yaramakala S, Margaritis D. Speculative Markov blanket discovery for optimal feature selection. Fifth IEEE International Conference on Data Mining (ICDM'05): IEEE. p. 809–12. https://www.ieee.org/publications/publications-contact.html.
29. Leinonen R, Sugawara H, Shumway M, et al. The sequence read archive. Nucleic Acids Res. 2011;39:D19–21.
30. Matias Rodrigues JF, Schmidt TSB, Tackmann J, et al. MAPseq: highly efficient k-mer search with confidence estimates, for rRNA sequence analysis. Bioinformatics. 2017. Epub ahead of print. https://doi.org/10.1093/bioinformatics/btx517.
31. Aliferis CF, Statnikov A, Tsamardinos I, et al. Local causal and markov blanket induction for causal discovery and feature selection for classification part ii: analysis and extensions. J Mach Learn Res. 2010;11:235–84.
32. Li J, Quinque D, Horz H-P, et al. Comparative analysis of the human saliva microbiome from different climate zones: Alaska, Germany, and Africa. BMC Microbiol. 2014;14:316.
33. Ravel J, Gajer P, Abdo Z, et al. Vaginal microbiome of reproductive-age women. Proc Natl Acad Sci U S A. 2011;108(Suppl 1):4680–7.
34. Qin J, Li R, Raes J, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature. 2010;464:59–65.
35. Segata N, Izard J, Waldron L, et al. Metagenomic biomarker discovery and explanation. Genome Biol. 2011;12:R60.
36. Gaci N, Borrel G, Tottey W, et al. Archaea and the human gut: new beginning of an old story. World J Gastroenterol. 2014;20:16062–78.
37. Hansen EE, Lozupone CA, Rey FE, et al. Pan-genome of the dominant human gut-associated archaeon, Methanobrevibacter smithii, studied in twins. Proc Natl Acad Sci U S A. 2011;108(Suppl 1):4599–606.
38. Mikx FH, Van der Hoeven JS. Symbiosis of Streptococcus mutans and Veillonella alcalescens in mixed continuous cultures. Arch Oral Biol. 1975;20:407–10.
39. Lloyd-Price J, Mahurkar A, Rahnavard G, et al. Strains, functions and dynamics in the expanded Human Microbiome Project. Nature. 2017;550:61–6.
40. Rinke C, Schwientek P, Sczyrba A, et al. Insights into the phylogeny and coding potential of microbial dark matter. Nature. 2013;499:431–7.
41. Sunagawa S, Mende DR, Zeller G, et al. Metagenomic species profiling using universal phylogenetic marker genes. Nat Methods. 2013;10:1196–9.
42. Young KD. The selective value of bacterial shape. Microbiol Mol Biol Rev. 2006;70:660–703.
43. Schulz HN, Jørgensen BB. Big Bacteria. Annu Rev Microbiol. 2001;55:105–37.
44. Hibbing ME, Fuqua C, Parsek MR, et al. Bacterial competition: surviving and thriving in the microbial jungle. Nat Rev Microbiol. 2010;8:15–25.
45. Stecher B, Barthel M, Schlumberger MC, et al. Motility allows S. Typhimurium to benefit from the mucosal defence. Cell Microbiol. 2008;10:1166–80.
46. Hyronimus B, Le Marrec C, Sassi AH, et al. Acid and bile tolerance of spore-forming lactic acid bacteria. Int J Food Microbiol. 2000;61:193–7.
47. Casula G, Cutting SM. Bacillus probiotics: spore germination in the gastrointestinal tract. Appl Environ Microbiol. 2002;68:2344–52.
48. O'Leary NA, Wright MW, Brister JR, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44:D733–45.
49. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
50. Chakravarti N. Isotonic median regression: a linear programming approach. Math Oper Res. 1989;14:303–8.
51. Woolf B. The log-likelihood ratio test (the G-test). Ann Hum Genet. 1957;21:397–409.
52. Sheridan C. The Python language reference manual: Lulu Press, Inc; 2016. https://www.lulu.com/.
53. Behnel S, Bradshaw R, Citro C, et al. Cython: the best of both worlds. Comput Sci Eng. 2011;13:31–9.
54. Benjamini Y, Hochberg AY. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol. 1995;57:289–300.
55. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics. 2013;29:2933–5.
56. Nawrocki EP. Structural RNA homology search and alignment using covariance models: Washington University in St. Louis; 2009. https://openscholarship.wustl.edu/etd/?utm_source=openscholarship.wustl.edu%2Fetd%2F256&utm_medium=PDF&utm_campaign=PDFCoverPages.
57. Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS One. 2010;5:e9490.
58. Wiki. MicrobeWiki. MicrobeWiki. https://microbewiki.kenyon.edu (2016).

Figure S1

a



b

RFC-global

RFC-single	U.P.	P.	Total
U.P.	17'017	15'863	32'881
P.	120	1'277	1'397
Total	17'138	17'140	34'278

c

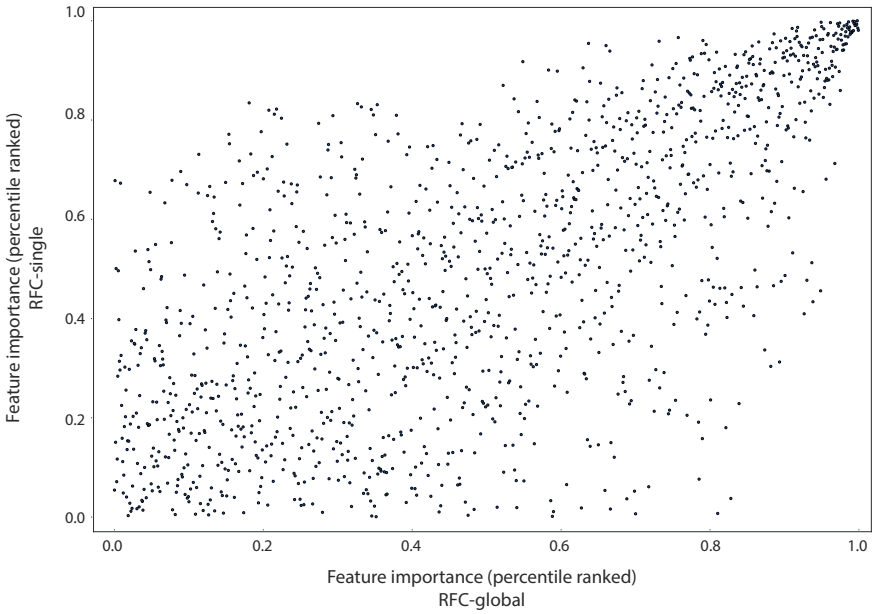


Figure S2

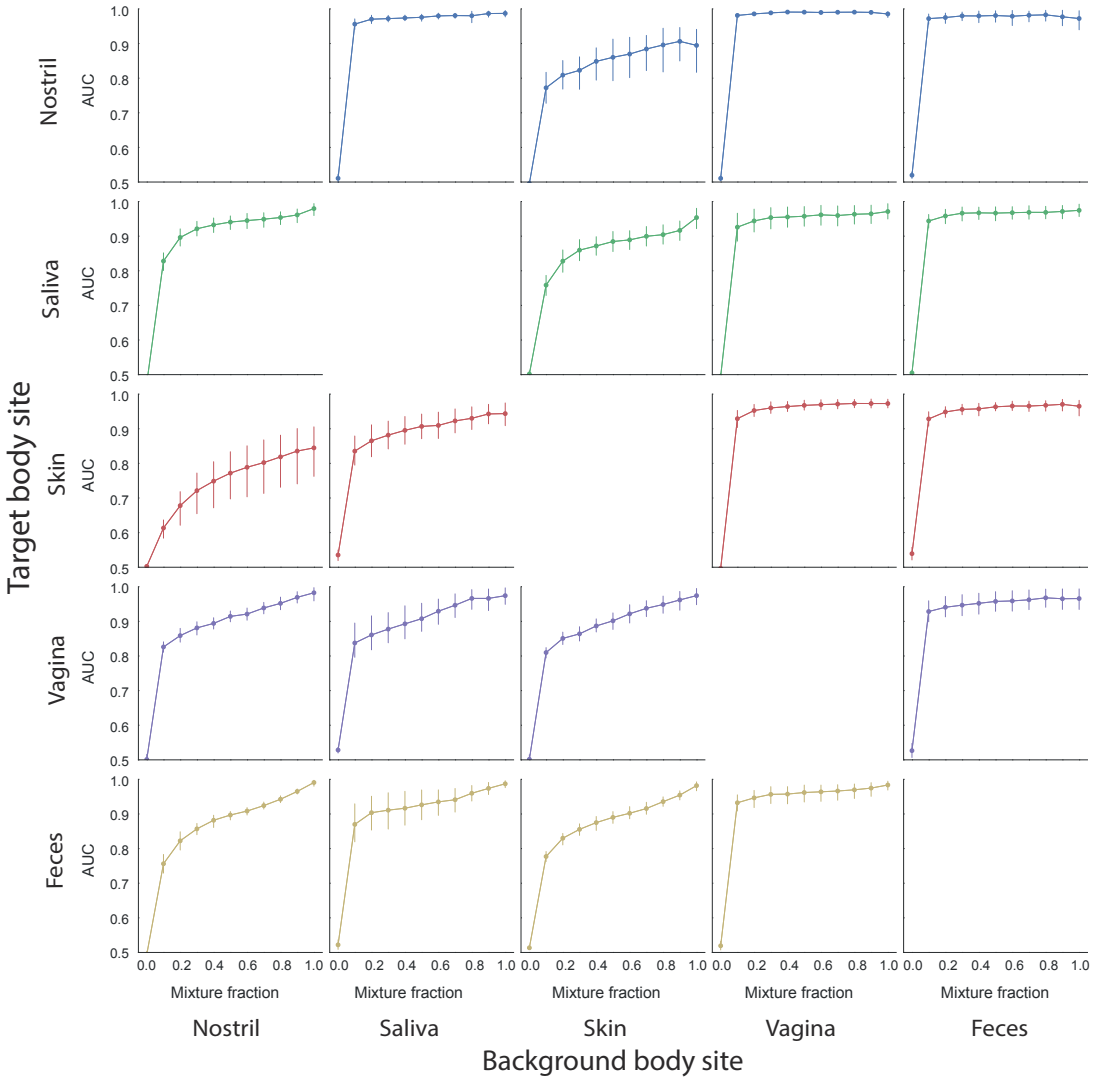

Target body site
Nostril
Saliva
Skin
Vagina
Feces
AUC
1.0
0.9
0.8
0.7
0.6
0.5
0.0
0.2
0.4
0.6
0.8
1.0
Mixture fraction
Nostril
Saliva
Skin
Vagina
Feces
Background body site

Figure S3

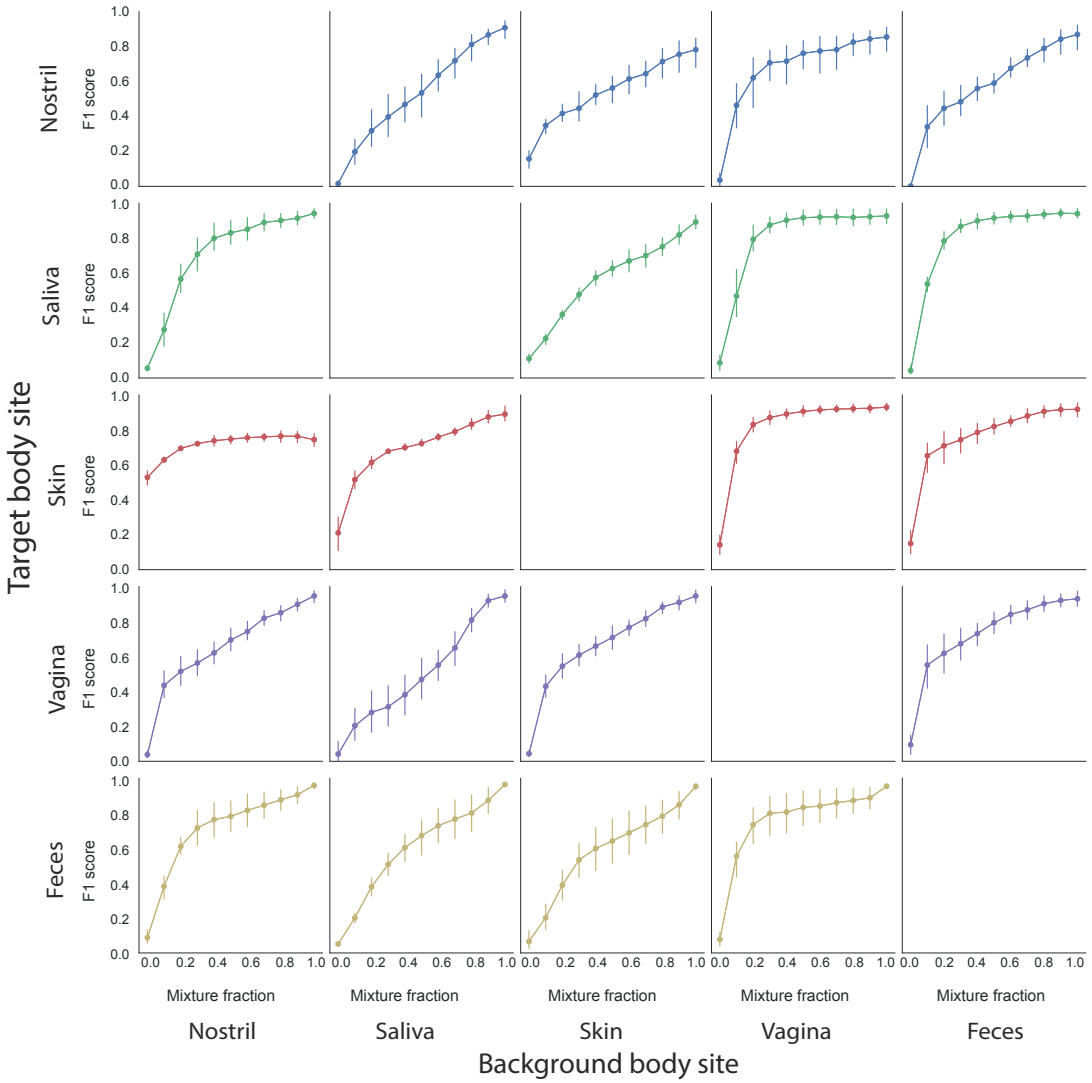

Target body site
Nostril
Saliva
Skin
Vagina
Feces
F1 score
0.0
0.2
0.4
0.6
0.8
1.0
Mixture fraction
Nostril
Saliva
Skin
Vagina
Feces
Background body site

Figure S4

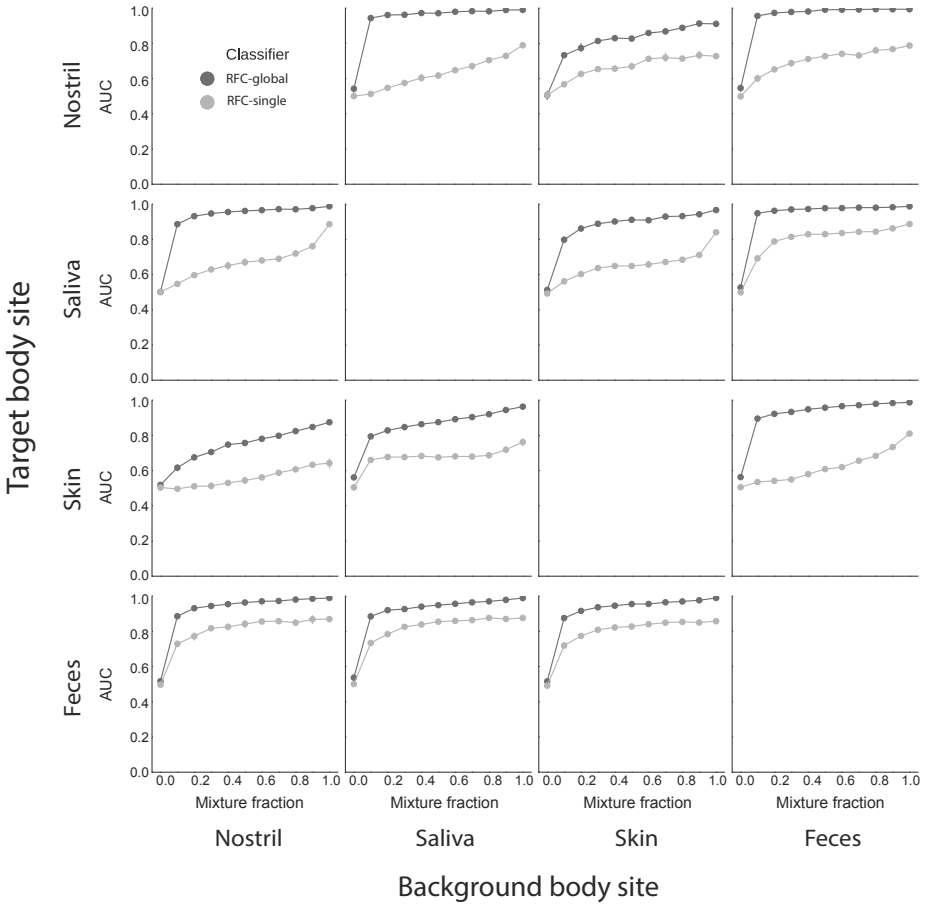

Classifier
RFC-global
RFC-single
Target body site
Nostril
Saliva
Skin
Feces
AUC
Mixture fraction
Background body site

Figure S5

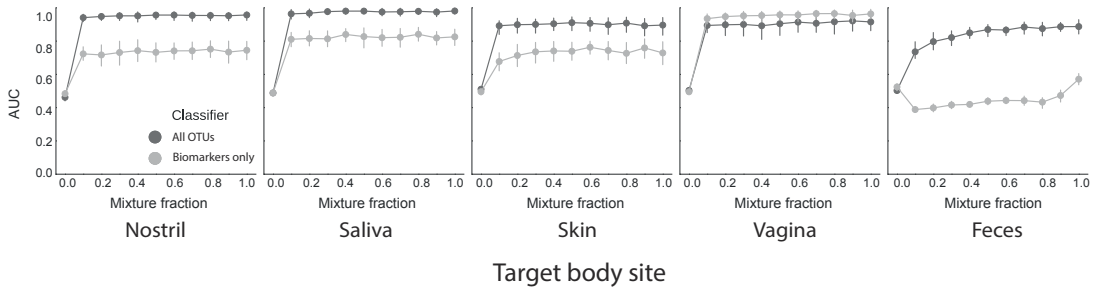


Figure S6

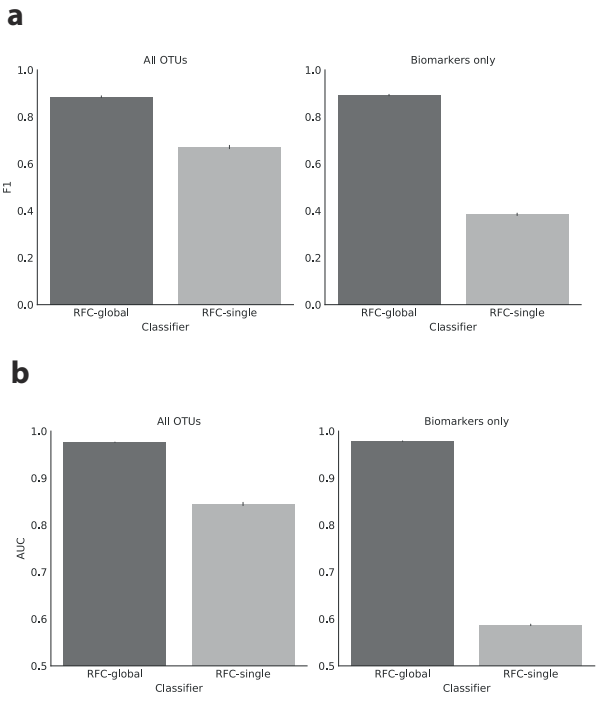


Figure S7

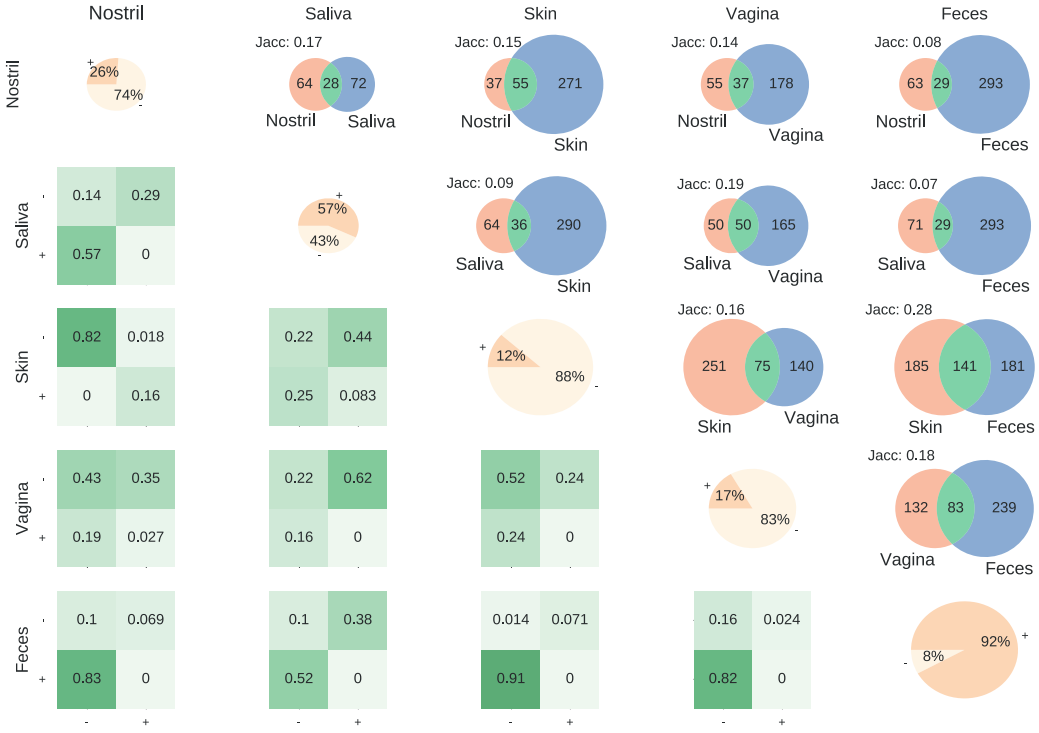


Figure S8

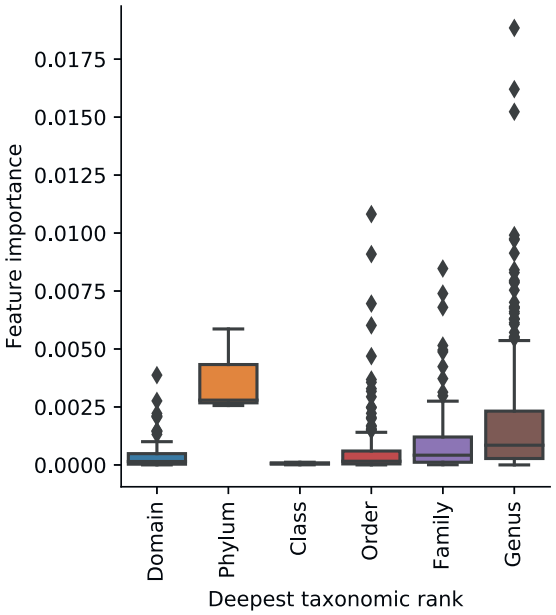


Figure S9

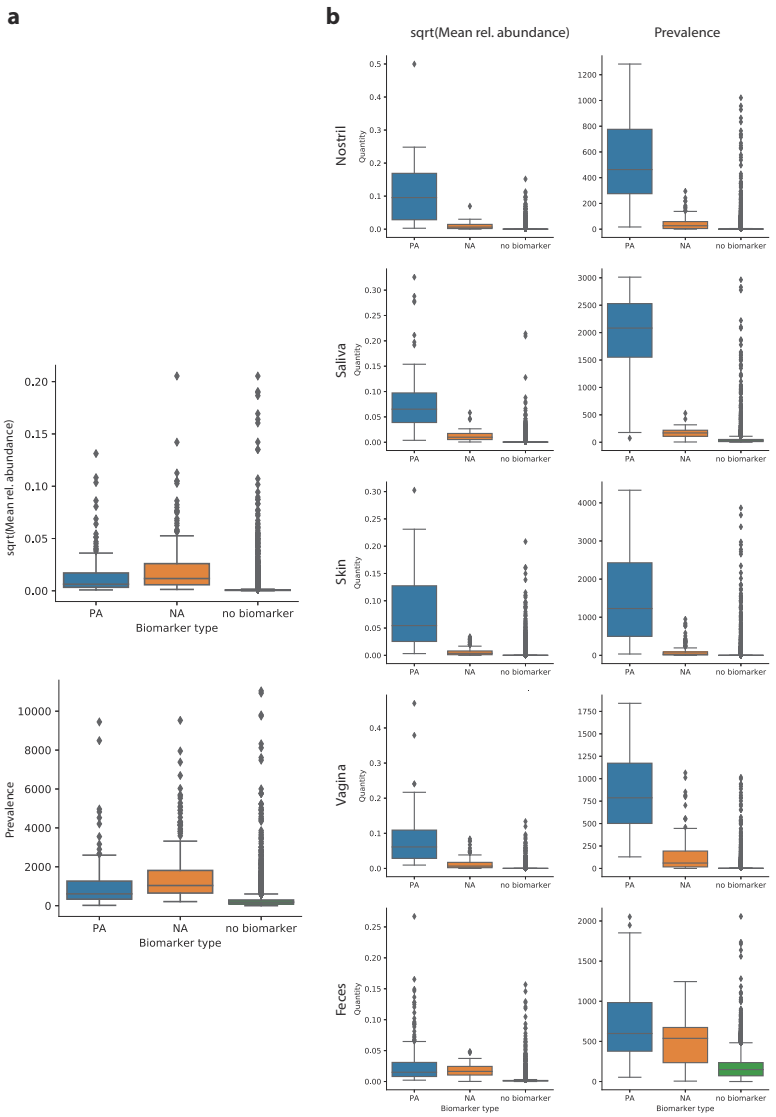

a
b
sqrt(Mean rel. abundance)
Prevalence
Nostril
Saliva
Skin
Vagina
Feces
Quantity
PA
NA
no biomarker
Biomarker type

Figure S10

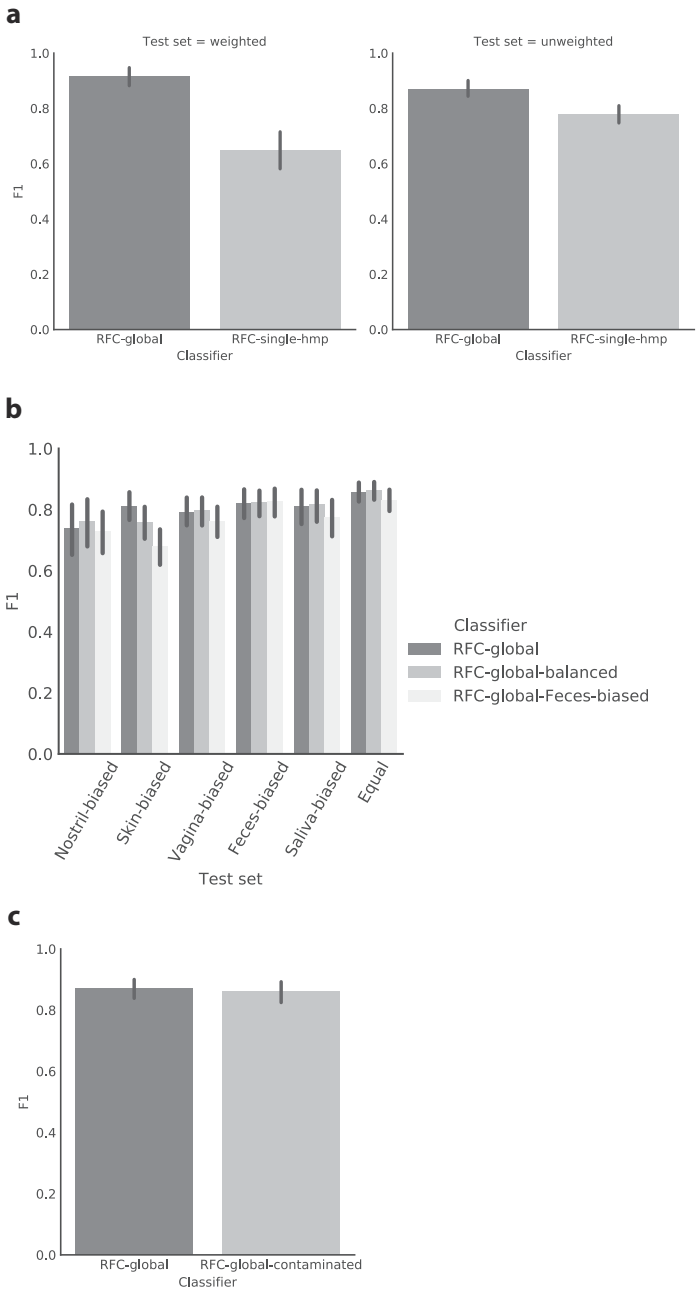

a
Test set = weighted
Test set = unweighted
F1
1.0
0.8
0.6
0.4
0.2
0.0
RFC-global
RFC-single-hmp
Classifier
b
F1
Nostril-biased
Skin-biased
Vagina-biased
Feces-biased
Saliva-biased
Equal
Test set
Classifier
RFC-global
RFC-global-balanced
RFC-global-Feces-biased
c
RFC-global
RFC-global-contaminated
Classifier

Figure S11

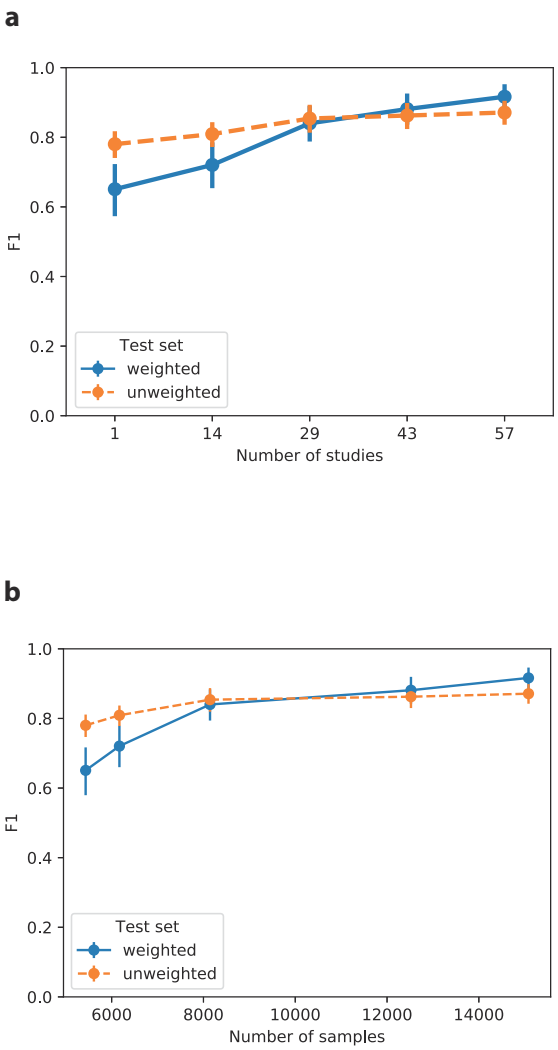

a
1.0
0.8
0.6
0.4
0.2
0.0
F1
Test set
weighted
unweighted
1
14
29
43
57
Number of studies
b
1.0
0.8
0.6
0.4
0.2
0.0
F1
Test set
weighted
unweighted
6000
8000
10000
12000
14000
Number of samples

Table S1

Body site	Genus	Number of positively associated OTUs	Percentile of Random Forest feature importance
Nostril	*Corynebacterium*	6	100.0
	Cutibacterium	2	99.8
	Dolosigranulum	1	93.2
	Acinetobacter	1	90.1
	Micrococcus	1	85.0
Saliva	*Veillonella*	2	99.3
	Prevotella	8	99.2
	Neisseria	1	98.9
	Actinomyces	1	98.1
	Gemella	2	98.0
Skin	*Corynebacterium*	2	100.0
	Cutibacterium	2	99.8
	Streptococcus	3	99.1
	Neisseria	1	98.9
	Gemella	1	98.0
Vagina	*Lactobacillus*	6	98.3
	Prevotella	3	95.7
	Finegoldia	1	94.6
	Gardnerella	1	94.5
	Atopobium	2	93.1
Feces	*Bacteroides*	21	93.9
	Enterococcus	2	88.8
	Parabacteroides	3	88.3
	Eubacterium	4	86.5
	Clostridium	5	82.2

Table S2

Body site	Genus	Number of positively associated OTUs	Reference
Nostril	Corynebacterium	6	[1]
	Cutibacterium	2	[2]
	Acinetobacter	1	[3]
	Delftia	1	not described in nostril
	Dolosigranulum	1	[4]
	Micrococcus	1	[4]
Saliva	Prevotella	8	[5]
	Selenomonas	2	[5]
	Capnocytophaga	4	[5]
	Actinomyces	1	[5]
	Gemella	2	[5]
	Leptotrichia	2	[5]
	Veillonella	2	[5]
	Fusobacterium	1	[5]
	Haemophilus	1	[5]
	Streptococcus	1	[5]
	Atopobium	1	[6]
	Rothia	1	[5]
	Peptostreptococcus	1	[6]
	Campylobacter	1	[5]
	Lautropia	1	[7]
	Catonella	1	[8]
	Granulicatella	1	[5]
	Porphyromonas	1	[5]
	Neisseria	1	[5]
	[Eubacterium] sulci	1	[8]
	Mogibacterium	1	[6]
	Lachnoanaerobaculum	1	[9]
Skin	Streptococcus	3	[11]
	Corynebacterium	2	[10]
	Paracoccus	1	[11]
	Gemella	1	[12]
	Sphingomonas	1	[11]
	Ralstonia	1	not described in skin
	Pseudomonas	1	[13]
	Moraxella	2	[12]
	Micrococcus	1	[13]
	Acinetobacter	1	[14]
	Actinomyces	1	[15]
	Delftia	1	[13]
	Acidovorax	1	[11]
	Cutibacterium	2	[2]
	Neisseria	1	[11]
	Caulobacter	1	not described in skin
	Fusobacterium	1	[12]
Vagina	Lactobacillus	6	[16]
	Anaerococcus	3	[17]
	Prevotella	3	[16]
	Atopobium	2	[16]
	Aerococcus	1	[16]
	Finegoldia	1	[16]
	Sneathia	1	[16]
	Mycoplasma	1	[18]
	Mycobacterium	1	[19]
	Actinomyces	1	[19]
	Gardnerella	1	[16]
	Ureaplasma	1	[18]
	Veillonella	1	[20]
Feces	Bacteroides	22	[21]
	Ruminiclostridium	1	[22]
	Clostridium	5	[21]
	Oscillibacter	1	[23]
	Eubacterium	4	[22]
	Alistipes	3	[24]
	Tyzzerella	2	[25]

Lachnoclostridium	5	[22]
Ruminococcus	3	[21]
Bifidobacterium	4	[21]
Faecalibacterium	1	[21]
Roseburia	1	[23]
Parabacteroides	3	[24]
Acidaminococcus	2	[22]
Collinsella	2	[25]
Anaerotruncus	1	[26]
Weissella	1	[27]
Anaerostipes	1	[28]
Bilophila	1	[29]
Blautia	5	[30]
Parasutterella	1	[24]
Odoribacter	1	[28]
Dorea	2	[23]
Lactococcus	1	[31]
Coprococcus	1	[22]
Akkermansia	1	[26]
Erysipelatoclostridium	3	[22]
Enterococcus	2	[21]
putative Halovenus	1	not described in gut
Holdemanella	2	[32]
Oxalobacter	1	[33]
Salmonella	1	[34]
Lactococcus	1	[31]
Lactobacillus	1	[21]
Microvirgula	1	[35]
Paraclostridium	1	[36]

References

[1] Frank DN, Feazel LM, Bessesen MT, et al. The human nasal microbiota and Staphylococcus aureus carriage. *PLoS One* 2010; 5: e10598.

[2] Fitz-Gibbon S, Tomida S, Chiu B-H, et al. Propionibacterium acnes strain populations in the human skin microbiome associated with acne. *J Invest Dermatol* 2013; 133: 2152–2160.

[3] Mahdavinia M, Keshavarzian A, Tobin MC, et al. A comprehensive review of the nasal microbiome in chronic rhinosinusitis (CRS). *Clin Exp Allergy* 2016; 46: 21–41.

[4] Kaspar U, Kriegeskorte A, Schubert T, et al. The culturome of the human nose habitats reveals individual bacterial fingerprint patterns. *Environ Microbiol* 2016; 18: 2130–2142.

[5] Li J, Quinque D, Horz H-P, et al. Comparative analysis of the human saliva microbiome from different climate zones: Alaska, Germany, and Africa. *BMC Microbiol* 2014; 14: 316.

[6] Gomes BPFA, Berber VB, Kokaras AS, et al. Microbiomes of Endodontic-Periodontal Lesions before and after Chemomechanical Preparation. *J Endod* 2015; 41: 1975–1984.

[7] Nasidze I, Li J, Quinque D, et al. Global diversity in the human salivary

microbiome. *Genome Res* 2009; 19: 636–643.

[8] Lazarevic V, Gaïa N, Girard M, et al. Comparison of DNA extraction methods in analysis of salivary bacterial communities. *PLoS One* 2013; 8: e67699.

[9] Johansson I, Witkowska E, Kaveh B, et al. The Microbiome in Populations with a Low and High Prevalence of Caries. *J Dent Res* 2016; 95: 80–86.

[10] Costello EK, Lauber CL, Hamady M, et al. Bacterial community variation in human body habitats across space and time. *Science* 2009; 326: 1694–1697.

[11] Cosseau C, Romano-Bertrand S, Duplan H, et al. Proteobacteria from the human skin microbiota: Species-level diversity and hypotheses. *One Health* 2016; 2: 33–41.

[12] Oh J, Freeman AF, NISC Comparative Sequencing Program, et al. The altered landscape of the human skin microbiome in patients with primary immunodeficiencies. *Genome Res* 2013; 23: 2103–2114.

[13] Grice EA, Kong HH, Renaud G, et al. A diversity profile of the human skin microbiota. *Genome Res* 2008; 18: 1043–1050.

[14] Smeekens SP, Huttenhower C, Riza A, et al. Skin microbiome imbalance in patients with STAT1/STAT3 defects impairs innate host defense responses. *J Innate Immun* 2014; 6: 253–262.

[15] van Rensburg JJ, Lin H, Gao X, et al. The Human Skin Microbiome Associates with the Outcome of and Is Influenced by Bacterial Infection. *MBio* 2015; 6: e01315–15.

[16] Ravel J, Gajer P, Abdo Z, et al. Vaginal microbiome of reproductive-age women. *Proc Natl Acad Sci U S A* 2011; 108 Suppl 1: 4680–4687.

[17] Zhou X, Brown CJ, Abdo Z, et al. Differences in the composition of vaginal microbial communities found in healthy Caucasian and black women. *ISME J* 2007; 1: 121–133.

[18] Marrazzo JM, Koutsky LA, Eschenbach DA, et al. Characterization of vaginal flora and bacterial vaginosis in women who have sex with women. *J Infect Dis* 2002; 185: 1307–1313.

[19] Huppert JS, Bates JR, Weber AF, et al. Abnormal vaginal pH and Mycoplasma genitalium infection. *J Pediatr Adolesc Gynecol* 2013; 26: 36–39.

[20] Redondo-Lopez V, Cook RL, Sobel JD. Emerging role of lactobacilli in the control and maintenance of the vaginal bacterial microflora. *Rev Infect Dis* 1990; 12: 856–872.

[21] Palmer C, Bik EM, DiGiulio DB, et al. Development of the human infant intestinal microbiota. *PLoS Biol* 2007; 5: e177.

[22] Creevey CJ, Kelly WJ, Henderson G, et al. Determining the culturability of the rumen bacterial microbiome. *Microb Biotechnol* 2014; 7: 467–479.

[23] Raman M, Ahmed I, Gillevet PM, et al. Fecal microbiome and volatile organic compound metabolome in obese humans with nonalcoholic fatty liver disease. *Clin*

Gastroenterol Hepatol 2013; 11: 868–75.e1–3.

[24] Shahinas D, Silverman M, Sittler T, et al. Toward an understanding of changes in diversity associated with fecal microbiome transplantation based on 16S rRNA gene deep sequencing. *MBio*; 3. Epub ahead of print 23 October 2012. DOI: 10.1128/mBio.00338-12.

[25] Joossens M, Huys G, Cnockaert M, et al. Dysbiosis of the faecal microbiota in patients with Crohn's disease and their unaffected relatives. *Gut* 2011; 60: 631–637.

[26] Biagi E, Nylund L, Candela M, et al. Through ageing, and beyond: gut microbiota and inflammatory status in seniors and centenarians. *PLoS One* 2010; 5: e10667.

[27] Walter J, Hertel C, Tannock GW, et al. Detection of Lactobacillus, Pediococcus, Leuconostoc, and Weissella species in human feces by using group-specific PCR primers and denaturing gradient gel electrophoresis. *Appl Environ Microbiol* 2001; 67: 2578–2585.

[28] Morgan XC, Tickle TL, Sokol H, et al. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol* 2012; 13: R79.

[29] Smith MI, Yatsunenko T, Manary MJ, et al. Gut microbiomes of Malawian twin pairs discordant for kwashiorkor. *Science* 2013; 339: 548–554.

[30] Rajilić–Stojanović M, Biagi E, Heilig HGHJ, et al. Global and Deep Molecular Analysis of Microbiota Signatures in Fecal Samples From Patients With Irritable Bowel Syndrome. *Gastroenterology* 2011; 141: 1792–1801.

[31] Barrett E, Kerr C, Murphy K, et al. The individual-specific and diverse nature of the preterm infant microbiota. *Arch Dis Child Fetal Neonatal Ed* 2013; 98: F334–40.

[32] Tanca A, Abbondio M, Palomba A, et al. Potential and active functions in the gut microbiota of a healthy human cohort. *Microbiome* 2017; 5: 79.

[33] Stewart CS, Duncan SH, Cave DR. Oxalobacter formigenes and its role in oxalate metabolism in the human gut. *FEMS Microbiol Lett* 2004; 230: 1–7.

[34] Ahmer BMM, Gunn JS. Interaction of Salmonella spp. with the Intestinal Microbiota. *Front Microbiol* 2011; 2: 101.

[35] Giamarellos-Bourboulis E, Tang J, Pyleris E, et al. Molecular assessment of differences in the duodenal microbiome in subjects with irritable bowel syndrome. *Scand J Gastroenterol* 2015; 50: 1076–1087.

[36] Tidjani Alou M, Million M, Traore SI, et al. Gut Bacteria Missing in Severe Acute Malnutrition, Can We Identify Potential Probiotics by Culturomics? *Front Microbiol* 2017; 8: 899.

Text S1

Supplementary Information

Optimal threshold estimation for the minimum number of OTUs per sample

We computed the pairwise unweighted Jaccard similarity between all samples to determine the closest neighbors of each sample in terms of OTU composition. Next, we calculated the entropy of body site labels amongst the top 100 closest neighbors of each sample in order to identify cases in which samples from multiple body sites show high similarity and are thereby hard to distinguish. We found a strong negative correlation between the body site entropy of closest neighbors and the number of unique OTUs per sample, which indicates that samples with low numbers of unique OTUs (usually due to low sequencing depth) from different body sites look similar in terms of OTU content. We identified the optimal threshold of unique OTUs at which this correlation disappeared (approximately 20 unique OTUs per sample) and applied this threshold to exclude noisy samples from our analysis.

Impact of unassignable reads on biomarker detection

OTUs assigned to *Staphylococcus*, a genus commonly reported as abundant skin inhabitant, could only be identified at comparatively low relative abundance and prevalence among skin samples in GlobalBodysites (5e-5 to 7e-5 abundance in 47 to 197 samples, calculated across all *Staphylococcus* OTUs univariately associated to skin). Upon further investigation, we found this to be caused by many reads confidently hitting this genus in the taxonomy database, but not mapping to a single 96% OTU in the reference database, but rather multiple OTUs with comparable alignment scores, making exact an OTU assignment impossible.

To evaluate the impact of these missing reads on biomarker detection, we created an abundance profile with reads that directly map to the *Staphylococcus* genus (0.24 average abundance across 3'247 skin samples) and applied GLL on this profile. We found a direct

association pattern between the *Staphylococcus* genus and skin, showing that despite the advantages of the 96% OTU definition (see discussion in the main text), it can lead to missing some biomarkers at more general taxonomic levels.

We also observed low relative abundance and prevalence for *Methanobrevibacter* in feces: OTUs assigned to this genus where found with lower abundance and prevalence than expected. Contrary to *Staphylococcus* however, the genus abundance profile analysis with GLL discarded *Methanobrevibacter* as direct biomarker, matching results from the OTU level analysis and showing that unassignable reads do not always confound biomarker detection.

As an additional note, we generally expect the usage of presence-absence normalization for biomarker detection as done in this study to reduce the effect of unassignable reads due to inherent robustness to noise in abundances.

CHAPTER 3
Discussion

3.1 Inference of microbial interaction networks

3.1.1 Improved network quality

In the first line of work, we developed and thoroughly evaluated FlashWeave, a new software framework for the inference of ecological microbial interaction networks based on statistical co-occurrence or co-avoidance in cross-sectional microbial abundance data (see Manuscript 2.1). In contrast to most other methods, FlashWeave aims at inferring highly interpretable and comprehensive networks, therefore its development necessitated special considerations in terms of direct ecological interactions, data heterogeneity, environmental and technical factors, as well as scalability. In this section, I will discuss the advances we made towards solving these issues in detail.

First of all, FlashWeave out-performed univariate co-occurrence methods in prediction performance benchmarks across diverse synthetic data sets, featuring both explicitly and implicitly modeled indirect associations. This reduced performance was due to consistently high numbers of false positive predictions for univariate methods, in line with extensive results from the PGM field (see subsection 1.3.3). Perhaps more interestingly, however: while alternative PGM approaches surpassed univariate methods, they nonetheless achieved generally lower prediction scores than FlashWeave in our benchmarks. A possible reason for this observation is that all PGM methods we tested infer direct edges by conditioning on all OTUs simultaneously, employing the L1 norm regularization objective (Lasso) to deal with high data dimensionality. In contrast, the incremental approach applied by FlashWeave includes only a subset of heuristically determined candidates into conditioning sets and furthermore imposes an upper bound on the conditioning set size. This more local approach to conditional independence search appears to have advantages over the Lasso on typical microbiome data sets, possibly because large sets of weakly relevant or irrelevant variables (Aliferis et al., 2010a) may result in power issues and more noisy estimates, both of which could be more effectively controlled through the strategy utilized by FlashWeave. Furthermore,

FlashWeave differs in its approach of handling compositional data (i.e. using adaptive instead of conventional pseudo-counts), and since some of the tested data sets show considerable compositionality, this adjustment may account in part for the observed improvements. However, these explanations are speculative and would require detailed validation.

Advantages for network reconstruction with FlashWeave also became apparent when inferring planktonic interaction networks from the TARA Oceans sequencing data set. When comparing the resulting networks to gold-standard planktonic interactions obtained from the literature, FlashWeave tended to rank these known interactions more highly within its inferred network than other tools. This effect was again particularly pronounced when comparing to univariate methods: these tools predicted substantially larger networks than PGM approaches, an observation that mirrored the results from our synthetic benchmarks and thus indicated large numbers of false positive predictions for univariate tools. As illustrated by Aliferis et al. (2010b), a phenomenon called "information synthesis" can result in indirect associations with higher weights than direct associations, which could possibly have lead to strong indirect associations with higher ranks than obtained for genuine interactions in univariate networks. However, in all likelihood, only a minority of true planktonic interactions has been described in the literature—definite conclusions on false positive predictions are thereby currently not possible.

The reduced performance of alternative PGM methods compared to FlashWeave on this real-world benchmark may again potentially be explained by the Lasso approach employed by these methods. Since in the Lasso workflow, edge weights are estimated based on conditioning on all other variables, weight predictions may be less optimal in the high-dimensional and sparse TARA Oceans data set than FlashWeave's local search. However, since not all true interactions are known, the possibility remains that some of these tools identified strong genuine interactions that were i) not described in the literature (or missed by the literature review) and ii) not detected by FlashWeave. It is noteworthy that all tested PGM-based approaches, including FlashWeave, share the conceptual similarity of approximating the true precision matrix of OTU-OTU interactions[1], making this scenario perhaps less likely. Nonetheless, more validated knowledge of planktonic ecosystems is required to rule out this possibility.

Particularly striking differences in network quality were observed for heterogeneous synthetic data sets composed of different disjoint habitats, which simulated a typical use case for cross-study analysis. In these benchmarks, all previously tested methods (including vanilla FlashWeave) reported greatly increased numbers of false positives compared to single-habitat benchmarks. However, the heterogeneity-aware mode of

[1] referring to the "sensitive" mode of FlashWeave (FlashWeave-S), which performed best in this benchmark

FlashWeave (termed FlashWeaveHE) achieved substantially improved performance over all other methods, predicting close to zero false positive edges at the cost of some reduction in sensitivity. The high precision was achieved through the specifically adjusted statistical tests employed by FlashWeaveHE, which account for excessive (i.e. structural or sampling) zeros, an important signature for niches and batches (Kaul et al., 2017). FlashWeaveHE also showed high consistency of inferred edges between single-habitat and multi-habitat data sets[2], for which we observed noticeable disparities in other approaches, further highlighting the robustness of FlashWeaveHE in scenarios featuring high sample heterogeneity.

The effective handling of such heterogeneous data sets is of utmost importance if the whole range of microbial community patterns (and consequently of inferred ecological interactions) is to be captured within sequencing data sets. Despite large-scale consortium efforts utilizing standardized and optimized experimental pipelines (Huttenhower et al., 2012; Thompson et al., 2017), no single protocol is ideal in all aspects. Ideally, a combination of approaches should be used to fully describe microbial communities (Sinha et al., 2017)—integrating information from several protocols into a single analysis is thereby enticing. Many confounders are furthermore known to be important microbiome modulators, with more being discovered constantly (Knight et al., 2018). Methods for handling unmeasured confounding effects could therefore be beneficial for any study, not just aggregated cross-study analyses. Nonetheless, the problem of unaccounted confounders is considerably aggravated in cross-study data sets, since these may span multiple habitats and be analyzed by different labs, which results in notoriously inconsistent and unreliable metadata annotations (Lagkouvardos et al., 2016; Mitchell et al., 2018; Pasolli et al., 2017). While standardized meta-annotation frameworks (Yilmaz et al., 2011) are not yet consistently used throughout all studies, methods accounting for unknown confounders are thus particularly crucial.

3.1.2 Scalability achievements

We found FlashWeave to be up to three orders of magnitude faster than other approaches—possibly more since some tools did not finish computation within our predetermined time frame. Interestingly, we also observed speed-advantages of FlashWeave compared to univariate methods, which tend to be conceptually more simple and typically incur smaller computational complexity. However, since the implementations of these tools could be less optimized than FlashWeave and may thus have a larger constant overhead, we expect differences to in principle vanish for larger data sets[3].

[2] the latter consisting of the same environments as the individual single-habitat data sets

[3] albeit we did not yet detect indications of this in our tests with increasing numbers of samples

As expected, Lasso-based PGM approaches were also much slower than FlashWeave, even though they typically rely on highly optimized solver software for linear systems. The most important explanation for this effect are algorithmic differences: the si-HITON-PC algorithm used by FlashWeave, in particular in conjunction with FlashWeave-specific heuristics, is highly efficient for sparse networks with large numbers of nodes, which is typical for microbial co-occurrence networks (Layeghifard et al., 2017; Röttjers and Faust, 2018). While hub nodes with many neighbors, also commonly observed in microbial co-occurrence networks (Banerjee et al., 2018; Faust and Raes, 2012), can pose problems to the si-HITON-PC algorithm (due to its exponential runtime dependence on neighborhood size), we found that the newly introduced heuristics in FlashWeave address this problem well in practice. In addition, edges involving at least one partner with few interactions, which make up the vast majority of all networks we tested, could be found very quickly in the networks we tested. As an option, an investigator may thus decide to start using this near-complete network structure for preliminary downstream analysis while FlashWeave is validating the final edges around large hubs, after which the initial results may be validated using the final structure.

Apart from algorithmic differences, another performance-critical distinction is FlashWeave's implementation with a heavy focus on large-scale input data. Our software is highly parallelized and runs seamlessly on High Performance Computing clusters, while also employing efficient data structures, such as sparse matrices (in tandem with specific algorithms to exploit their special properties), to minimize runtime and memory footprints. Furthermore, the newly introduced adaptive pseudo-counts used by FlashWeave resulted in large speed-ups compared to conventional pseudo-counts in our benchmarks[4]. This was achieved by reducing spurious edges involving rare OTUs, which can result in considerably decreased runtimes for the conditional independence search step performed by FlashWeave.

In addition, the ability of FlashWeaveHE to discard large numbers of spurious associations in highly heterogeneous data sets resulted in marked speed-ups on such data compared to all other methods, including vanilla FlashWeave. We demonstrated FlashWeaveHE's scalability on a heterogeneous real-world data set by learning a global, cross-habitat network based on more than half a million sequencing samples and more than seventy thousand OTUs. FlashWeaveHE fully inferred this network in less than two days with moderate computational resources, demonstrating its ability to readily analyze even the largest modern data sets.

As a finishing note, FlashWeave's performance is highly tunable: many parameters and options can be used to further alter the tradeoff between speed and network quality. For instance, FlashWeave can be asked to not use the full quantitative association

[4]Again, note that this normalization scheme is restricted to the "sensitive" mode of FlashWeave (FlashWeave-S).

information, but instead discretize data prior to analysis. This option (termed the "fast" mode, as opposed to the "sensitive" mode) yields a more coarse-grained view on microbial co-occurrence patterns and generally results in less sensitivity, but in exchange may improve runtime. This is due to the typically increased sparsity of these coarse-grained networks, which leads to fewer conditional independence tests. Similarly, other parameters related to specific heuristics, for instance the maximum conditioning set size and several reliability thresholds for statistical tests, can be adjusted—with potentially drastic impact on network inference speed.

3.1.3 Biological insights

We analyzed an aggregated cross-study data set of almost seventy thousand publicly available human gut samples (which we refer to as the "Global Gut" data set) and found a number of previously described biological patterns which, albeit detectable in both the univariate and the conditional networks, were more striking in the latter.

For instance, the phylogenetic assortativity signal of positive edges was considerably more pronounced in the conditional network, a fact not easily explainable by niche effects, since these are expected to be stronger in the univariate network (see subsection 1.3.3). While we suggest kin selection as a possible mechanism to create this pattern, this hypothesis is speculative and more in-depth investigation is required to further elucidate this point. This may for instance be pursued through comparative genomic analyses of interaction partners in combination with predicted physiological traits, which could allow to pinpoint possible mechanisms and to devise evolutionarily plausible scenarios.

We furthermore found that positive interaction patterns between H_2 producers and consumers were more pronounced when removing indirect edges, even when accounting for confounding phylogenetic relatedness (see the previous paragraph). Since these groups of organisms are known to form intimate syntrophic relationships in the human gastrointestinal tract and have been subject to extensive research (Carbonero et al., 2012), the observed signal is potentially biologically meaningful. Furthermore, marked advantages of removing indirect edges are plausible and even expected in this sub-system, since it is rich in shared ecological dependencies, which are explicitly targeted by our approach. This finding thus clearly illustrates how PGM approaches that remove such indirect associations can improve the interpretability of predicted networks.

Underlying these observed differences in biological interpretability is the fact that the vast majority of univariate edges (96%) were discarded as indirect by FlashWeave's conditional independence search. While the set of true interactions in this data set is unknown, this result nonetheless is consistent with our previous results based on

synthetic data sets and thus demonstrates the impact of our PGM approach on model sparsity and, in consequence, interpretability.

Further patterns of biological interest in the Global Gut network include positive and negative hub OTUs, which have particularly high numbers of directly positively and negatively associated neighbors, respectively. While negative hubs featured a number of species associated with gastrointestinal inflammation and dysbiosis in the literature, many positive hubs were classified as taxa known to harbor important mutualist species that benefit gut health. Strikingly, the most dominant positive hubs also included several OTUs that were only classifiable at the family level and thereby constitute candidates for novel keystone taxa in the human gastrointestinal tract. Further confirmatory research on these candidates, for instance through closer examination of the studies they were found in and detailed taxonomic and genomic investigation of positive interaction partners, has the potential to provide valuable insights into gut ecosystem homeostasis, defense mechanisms and integrity.

Rare species are typically removed in co-occurrence network analysis, typically justified by considerations regarding the detection limit of utilized sequencing technologies, which may induce significant numbers of sampling zeros in less abundant species (Banerjee et al., 2018; Röttjers and Faust, 2018). This results in overall lower count information for these species, making the detection of genuine interaction patterns less likely, while also inducing spurious positive associations between rare species (and negative associations between rare and abundant species). Nonetheless, less abundant species are known to play important auxiliary roles in microbial ecosystems and even constitute keystone taxa (Banerjee et al., 2018; Jousset et al., 2017; Lynch and Neufeld, 2015; Sogin et al., 2006). We thus aim to be more inclusive, now that modern cross-study data sets provide sufficient statistical power to support analyses of rare taxa, and furthermore aided by specific optimizations in FlashWeaveHE to account for sampling zeros.

The effect of this more inclusive approach was clearly visible in the Global Gut data set: a network based on only the single largest study (the American Gut Project) within this data set missed the majority of overall edges compared to the full network. Notably, the majority of undetected edges featured at least one rare OTU with insufficient prevalence in the American Gut Project. This initial glimpse at massive, previously underappreciated association patterns in the rare ecosphere highlights the need for more focused investigation of the long tail of rare microbes (Jousset et al., 2017; Lynch and Neufeld, 2015).

3.1.4 Limitations and outlook

While heterogeneity alters presence-absence patterns of OTUs, as captured by structural zeros and thus mitigated by FlashWeaveHE, it may also affect relative abundances. So far, no FlashWeaveHE-specific optimizations have been introduced to handle this effect. The first reasoning behind this choice is that we expect structural zeros to be the dominant source of spurious associations in highly heterogeneous data sets, since microbial species are known to stratify considerably according to habitats and conditions (Knight et al., 2018; Röttjers and Faust, 2018; Thompson et al., 2017). In this setting, every additional study is expected to add rapidly increasing numbers of structural zeros to the OTU table, partly due previously observed OTUs not being present in the new study, but mostly due to newly introduced OTUs generating structural zeros in all previously added data sets. This quickly results in the majority of entries in the OTU table constituting structural zeros, a pattern we empirically find in most heterogeneous data sets we tested. We thus expect that accounting for these zeros addresses the most critical source of spurious edges.

The second reason for not specifically considering heterogeneity-induced abundance changes is that we empirically observe unmeasured confounders to introduce concerted abundance shifts in many OTUs at once. FlashWeave is often able to detect this shared signal in our tests—even if the confounding variable was not explicitly provided as an input—and thus tends to remove the corresponding spurious associations. For instance, when analyzing the Human Microbiome Project (HMP) data set, we observed that most associations discarded through primer or body site meta variables were also discarded when these variables were not available to FlashWeave. Similarly, when simulating dependent sample groups, representing for instance samples from the same subject in a time series or multiple sequencing runs of the same sample (generated using different experimental protocols), the conditional independence search discarded large fractions of spurious associations introduced through these confounders, even though explicit information on group membership was not provided to FlashWeave.

These observations can be explained by OTUs strongly associated with unmeasured confounders taking their place in conditioning sets. Albeit the OTUs are only an approximation to the underlying latent variable, they nonetheless enable the removal of many spurious edges normally discarded by directly conditioning on this confounder. However, as clearly demonstrated in our benchmarks on synthetic heterogeneous data, this effect has limits and thus other means, such as the handling of structural zeros in FlashWeaveHE, are necessary for accurate network recovery. Similarly, while the global structure in the HMP network was only marginally affected when omitting meta variable information, hotspots of spurious associations within the neighborhoods fo these variables still were apparent. This observation highlights that explicit meta

information should still be provided to FlashWeave whenever available. A possible avenue for future improvements is the more direct handling of unknown confounders through ideas from the fast causal inference (FCI, Spirtes et al. (2000)) algorithm, which explicitly handles latent variables, but has a higher computational complexity.

Another limitation in the context of heterogeneity is that the handling of structural and sampling zeros by FlashWeaveHE is currently done conservatively by discarding all absences, which potentially also includes ecologically informative zeros. While we found the reduction in sensitivity introduced by this choice to be modest in our benchmarks, especially for the large data sets that FlashWeaveHE was designed for, it may still be worthwhile to explore more complex statistical approaches in the future to probabilistically distinguish excess zeros from informative zeros and thereby increase statistical power. A variety of methods, including zero-inflated statistical models and other approaches, have recently been proposed to this end (Chen and Li, 2016; Kaul et al., 2017) and may inspire future versions of FlashWeave. However, these approaches incur increased computational costs and whether they can handle the size of modern cross-study data sets is yet to be determined.

A critical aspect of the GLL-LGL framework used by FlashWeave is that it yields causal inference algorithms. While FlashWeave currently only implements the skeleton learning step of this framework, and thus predicts undirected networks, follow-up procedures to infer (partial) edge directionality are an integral part of GLL-LGL algorithms and could be added seamlessly (see subsection 1.3.3). However, these steps can be computationally expensive and would likely require careful implementations, as well as possibly the invention of additional heuristics similar to those needed for FlashWeave's skeleton learning algorithm. Since the predicted graphs would only be partially directed, Functional Causal Models (FCMs), which can predict edge directionality more reliably through additional assumptions, may also be considered in the future (Zhang et al., 2018a). Furthermore, Structural Equation Models (SEMs) could be used to allow for directed cycles in predicted networks, which are not allowed in Bayesian Networks (Spirtes, 2010). Interestingly, causal inference has recently been suggested as a particularly suited framework for analyzing heterogeneous data sets in a principled fashion[5] (Zhang et al., 2017; Zhang et al., 2018a), which makes the inclusion of causal inference steps into future versions of FlashWeave also important in the context of improved data heterogeneity handling.

Another interesting implication of directionality inference is the learning of full predictive models: as soon as all edges are directed, the full parameters of the model's joint probability distribution can be efficiently learned, which allows predictions of individual abundances and may for instance be interesting to forecast the effect of

[5]Conversely, data heterogeneity can additionally inform directionality prediction (Zhang et al., 2018a).

perturbations on the system. If a fully directed network is not available, the partially directed graph produced by simple edge-directing rules can nonetheless be used as an input to search-and-score or MCMC sampling algorithms. However, while this may allow for more tractable learning of the full model, such approaches for learning full Bayesian Networks remain notoriously expensive (NP-Hard).

Nonetheless, rapid advances in variational inference and MCMC sampling techniques (Blei et al., 2017; Robert et al., 2018), such as Hamiltonian Monte Carlo methods (Neal et al., 2011), are already enabling new types of sampling software that provide substantial leaps in computational efficiency for broad applications (Hoffman and Gelman, 2014). This idea was recently pushed to its extreme by Regier et al. (2018), who inferred a full generational model of the visible universe, with parameters for 188 million stars, from 55 terabytes of telescope image data. This model was learned using a thoroughly optimized Julia implementation of a variational inference algorithm (called Celeste), run on a high-end petascale supercomputer, and represents one of the largest graphical models available in any domain to date. While invested effort and resources were immense, this milestone of graphical model inference nonetheless provides exciting prospects for the next generation of ecological modeling efforts, which could benefit from the currently undergoing data accumulation in the microbiome field[6].

Another emerging topic are specialized hardware architectures, such as graphics processing units (GPUs) or, more recently, tensor processing units (TPUs) and field-programmable gate arrays (FGPAs) (HajiRassouliha et al., 2018). While these specialized processors are already extensively used to boost numerical computation, in particular machine learning applications in other research fields and industry (Abadi et al., 2016), they are so far underutilized in microbial ecology. Fortunately, the Julia programming language (Bezanson et al., 2017), in which FlashWeave is implemented, readily embraces new types of hardware and thus provides for instance generic and maintainable ways to compile Julia programs to run directly on GPUs (Besard et al., 2018). Similar principles have furthermore lead lately to tremendous speed-ups of Julia programs through direct cross-compilation to TPUs (Fischer and Saba, 2018). While FlashWeave currently faces no discernable computational limits[7], sequence databases are still growing exponentially and computational burden thus keeps increasing. Ready integration of modern hardware types is thereby a crucial property of any software to ensure its future-proofness.

Many recent microbiome data sets, in particular those covering nation-scale human cohorts in the scope of health-related studies (Falony et al., 2016; McDonald et al., 2018;

[6] in terms of raw sequencing data, quantities are indeed of an even larger scale than the data fueling Celeste (see subsection 1.2.4 and citations therein)

[7] as previously mentioned, it learned a network from more than half a million samples and tens of thousands of OTUs in less than two days using moderate resources

Zhernakova et al., 2016), feature detailed metadata information on various factors, for instance variables related to lifestyle and health status. But also environmental data sets are becoming more metadata rich and now cover increasingly broad physicochemical variables, such as pH, temperature or salinity (Sunagawa et al., 2015; Thompson et al., 2017). These detailed data sets create a unique opportunity for integrative ecosystem analysis that combines microbial and non-microbial variables. Explicit inclusion of meta variables into ecosystem models can reveal direct associations between these variables and individual microbial species. These associations could, for instance, shed light on how environmental conditions influence important ecosystem members and how this influence may radiate throughout the whole system. We thus far quantified the importance of meta variables only within the HMP network, but more detailed analysis of other metadata rich microbiome data sets would be an intriguing endeavor. This may help disentangle complex relationships within these intricate systems and potentially allow the discovery of new direct links between microbes and meta variables, such as diseases or lifestyle variables.

Another important aspect of FlashWeave is that its integral statistical tests and algorithms are generic and thus in principle applicable to other data types. Indeed, related algorithms have been used in a wide variety of fields (see subsection 1.3.3) and not all novelties introduced in FlashWeave are necessarily specific to microbial abundance data. For instance, the heuristics to accelerate computations on hubs are likely to be relevant in gene or protein interaction networks, which are also known for scale-free node degree distributions and the presence of hubs (Luscombe et al., 2004; Tsai et al., 2009). Similarly, co-occurrence network analysis of traditional animal-plant ecosystems is generally possible in FlashWeave (albeit so far untested) and may become an interesting prospect in the face of increasingly available cross-sectional data on these ecosystems. Furthermore, the adaptive pseudo-count scheme devised for FlashWeave may have applications in other areas of compositional data analysis, which could be explored in the future.

An interesting alternative use-case for FlashWeave would be integrative multi-omics data analysis, in which diverse types of data are combined into a single data set and correlation structures tend to be complex, making interpretation challenging (Hasin et al., 2017; Heintz-Buschart and Wilmes, 2018). Since FlashWeave distinguishes direct and indirect effects between all variable types (not restricted to OTUs or genes, for instance), it could potentially be used to draw a more comprehensive and interpretable picture of how different variables, such as OTUs, functional genes and metabolites, are related to each other and target variables of interest.

3.2 Ecologically informed biomarker discovery

3.2.1 Conceptual advantages

As mentioned previously, the detection of interpretable biomarkers is a key goal of modern microbiome research (see subsection 1.2.5). Advances in biomarker detection from microbiome data have recently been made to account for compositionality and excess zeros (Kaul et al., 2017; Mandal et al., 2015), or to discover more complex, nonlinear association patterns via supervised learning algorithms (Knights et al., 2011a; Pasolli et al., 2016; Statnikov et al., 2013b). Yet, the confounding effect of ecological interactions between microbial species in driving associations between species and variables of interest is typically not considered.

Causal inference approaches are prime candidates for principled, ecologically informed biomarker discovery, because they are "optimal" in the sense of discovering directly associated links and removing spurious (i.e. purely correlational) associations that are typically reported by other classes of methods (Aliferis et al., 2010a; Aliferis et al., 2010b). As I illustrated in this thesis, ecologically informed biomarker discovery and ecological network construction with meta variables can be seen as two sides of the same coin: while meta variables can drive spurious association patterns between microbial species and should therefore be accounted for in network construction, interactions between microbial species can similarly introduce spurious associations between microbial species and variables of interest in the context of biomarker discovery. Thus, the same set of algorithms used by FlashWeave for network learning can be used for detecting ecologically informed biomarkers when restricted to the local learning step, which yields a theoretically sound feature selection method (Aliferis et al., 2010a).

The prime advantage of this approach is that it only reports associations that are independent of any other variable in the system and thus represent particularly interesting relationships. For instance, biomarkers directly associated with certain habitats (such as human body sites) would be expected to be intimately tied to their preferred habitat by fundamental physiological constraints, rather than ecological dependencies on other species. Such physiological constraints, as predicted by direct associations, could subsequently be unraveled in targeted analyses. Similarly, direct disease associations present more promising candidates for causative agents, since species indirectly associated with these diseases are more likely to be part of a dysbiotic response to the altered environment (induced by the disease state) rather than being causally involved. While unmeasured variables may still turn out to be more ultimate causes for a disease than directly associated species in such cases, the biomarkers detected by this approach nonetheless provide a closer approximation to these unknown factors and may thus effectively guide further investigation towards them.

3.2.2 Application to human body site microbiota

The framework on which FlashWeave's local learning step is based (GLL) was shown to provide a variety of advantages for feature selection across different types of data (Aliferis et al., 2010a; Aliferis et al., 2010b) and also displayed superior parsimony for detected microbial biomarkers of a skin disease (Statnikov et al., 2013a). We thus decided to test its capabilities on the problem of biomarker detection for five human body sites, based on a heterogeneous, large-scale data set with more than fifteen thousand samples from over fifty studies. This prediction problem is particularly interesting in the context of accurate forensic body fluid identification and environmental contamination monitoring, but also more generally to gain a more precise understanding of direct body site-specific association patterns for individual microbial species. Notably, no ecologically informed biomarker selection method had previously been applied to human body sites to our knowledge, in particular not at the level of heterogeneity and scale employed in our study. This provided us with the unique opportunity to infer and report a more general and compact core set of body site-specific microbial biomarkers, the properties of which we could investigate in detail, and to furthermore test the predictive power of this biomarker set in a challenging evaluation across highly heterogeneous studies.

Our biomarker selection approach reduced the data set from tens of thousands of input OTUs to a condensed set of several hundred biomarker OTUs, which were directly positively or negatively associated with at least one body site. In line with Statnikov et al. (2013a), we found the selected biomarkers to be highly predictive when used to train a supervised classifier, achieving strongly improved performance compared to classifiers based on biomarkers and samples from single studies. Crucially, this performance also extended to mixtures of different body sites, where even small fractions of microbial communities from one body site could be accurately detected within large fractions of microbial communities from other body sites. This remarkable predictive performance on both mixed and unmixed communities across highly heterogeneous human samples showcases the relevance and generality of the discovered biomarkers. Interestingly, negative biomarkers showed higher predictive performance than positive biomarkers for this classification problem, indicating that absence information can be crucial for accurate predictions, a fact that appears to be rarely considered in the current microbiome literature.

When delving deeper into positive and negative association patterns between biomarker OTUs and body sites, we found that a number of associations reported in the literature were labeled as indirect in our analysis. We explored selected cases in more detail and indeed found known ecological dependencies of these discarded OTUs to other, directly associated taxa in the literature, indicating the relevance of

our ecologically informed approach. We furthermore detected several genus-body site associations that to our knowledge have not been previously described in the literature and additionally identified a noticeable fraction of so-called microbial dark matter among the most predictive biomarkers. The latter is intriguing because it indicates that there could still be considerable uncertainty in terms of body site-specific microbes, despite the extensive research that has been conducted on human-associated microbiota in recent years. It thus seems, that large, heterogeneous data sets in combination with novel biomarker prediction methods can enable fresh insights, even into widely studied ecosystems.

3.2.3 Limitations and outlook

While the GLL approach has been thoroughly validated in prior studies (Aliferis et al., 2010a; Aliferis et al., 2010b), evaluations on microbiome data are to our knowledge currently restricted to a study conducted by Statnikov et al. (2013a) on a human skin disease and our application of FlashWeave to the problem of biomarker discovery for human body sites. While these studies yielded promising results, more detailed biomarker detection benchmarks for microbiome data would thereby be desirable. Ideally, these would be performed on a variety of different microbial communities in a synthetic setup with known associations between species and target variables, with explicitly modeled ecological associations between microbes, but also featuring different types of target variables and data set sizes. Such a setup would allow the more proper assessment of advantages and limitations of FlashWeave (and other methods) in the context of biomarker discovery under confounding ecological interactions.

In particular, it is known that sensitivity can be decreased in tests for conditional independence, necessitating generally larger data sets than would be required for univariate methods (Aliferis et al., 2010a), but exact limits are so far unclear in the context of microbiome data. Should power issues be detected in the suggested benchmarks, it would be interesting to explore alternative parametric tests with higher sensitivity, for instance based on zero-inflated negative binomial regression (Xu et al., 2015), which could be plugged into FlashWeave instead of the currently used partial correlation or mutual information tests. However, as in the case of ecological network inference, we expect the scale of currently available microbiome data to provide sufficient statistical power for ecologically informed biomarker discovery in cross-study settings.

While we found indications of improved biomarker quality through our ecologically informed approach, the scarcity of experimentally observed interactions still hampers full verification of the direct associations we report. While we performed an initial exploratory analysis of genus-level physiological traits associated with our predicted

biomarker OTUs (based on literature), which shed some light on traits potentially driving the observed direct association patterns between microbes and body sites, this analysis could be extended to a comparative genomics level. This may allow the detection of more numerous and detailed traits, such as pathways and important functional gene categories, which may unveil more subtle body site-specific adaptations within the selected biomarkers.

Furthermore, many OTUs were not taxonomically classified at sufficient depth for detailed follow-up analysis, thus assembling genomes for some of these highly predictive OTUs in order to enable detailed study of their body site preferences would be an intriguing prospect. Future adoption of the employed workflow to other microbiome-related problems, such as disease prediction and other demanding biomarker and classification tasks, would be another interesting way forward.

Finally, the outstanding prediction performance and robustness of the classifier trained in this study indicate the potential of such methods for eventual utilization in court cases involving body fluid evidence. However, in-depth validation studies, as well as a detailed understanding of the employed classifiers (which are typically "black boxes"), would be an essential precondition.

3.3 Concluding remarks

We now live in an age of bountiful, globally distributed sequencing data, covering a plethora of environments, sampled under various conditions and analyzed with diverse protocols. This wealth of information allows unprecedented insights into the smallest, yet most important, inhabitants of our planet: microbes. In these times of rich and complex data, also the importance of heterogeneity-aware and scalable data analysis is increasingly recognized—not only in microbial ecology (Dai et al., 2018; Röttjers and Faust, 2018), but also other fields of study (Leek and Storey, 2007; Zhang et al., 2018a). While many components constituting biological systems, including microbial ecosystems, have now been identified and described in increasing detail, a new wave of holistic thinking, with the aim to integrate these parts into comprehensive models of reality, is rapidly gaining traction ("Systems biology", Chuang et al. (2010)).

As a contribution to these emergent developments, we here presented FlashWeave: a highly optimized framework, inspired by causal inference, for learning comprehensive and interpretable ecological networks and biomarkers from modern, large-scale cross-study data sets. The strong performance we see in a variety of benchmarks, complemented by an encouraging consistency with prior ecological knowledge, as well as novel patterns we detected in one of the largest reconstructed gut ecosystem models to date, make us confident that FlashWeave can be an important factor in pushing interpretable microbial ecosystem modeling to the next level. It has the potential to substantially improve our understanding of emergent properties of microbial ecosystems and to catalyze diverse applications, such as pro- and antibiotic development, next-generation culturing efforts and, in the future, ecosystem engineering. Much is yet to be learned about the immense, interconnected ecosystems spanning our planet, but the lense of network science may guide the way.

Appendix A

MAPseq: highly efficient k-mer search with confidence estimates, for rRNA sequence analysis

Contribution: Janko Tackmann (JT) contributed to testing the software and performed the analysis and visualization for the HMP primer consistency benchmark (Fig 1 g,h in the manuscript). JT furthermore contributed to reviewing and editing of the final manuscript.

Bioinformatics, 33(23), 2017, 3808–3810
doi: 10.1093/bioinformatics/btx517
Advance Access Publication Date: 14 August 2017
Applications Note

Sequence analysis

MAPseq: highly efficient k-mer search with confidence estimates, for rRNA sequence analysis

João F. Matias Rodrigues, Thomas S. B. Schmidt, Janko Tackmann and Christian von Mering*

Department of Molecular Life Sciences, and Swiss Institute of Bioinformatics, University of Zurich, Zurich, Switzerland

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on April 13, 2017; revised on July 19, 2017; editorial decision on August 8, 2017; accepted on August 10, 2017

Abstract

Motivation: Ribosomal RNA profiling has become crucial to studying microbial communities, but meaningful taxonomic analysis and inter-comparison of such data are still hampered by technical limitations, between-study design variability and inconsistencies between taxonomies used.
Results: Here we present MAPseq, a framework for reference-based rRNA sequence analysis that is up to 30% more accurate ($F_{1/2}$ score) and up to one hundred times faster than existing solutions, providing in a single run multiple taxonomy classifications and hierarchical operational taxonomic unit mappings, for rRNA sequences in both amplicon and shotgun sequencing strategies, and for datasets of virtually any size.
Availability and implementation: Source code and binaries are freely available at https://github.com/jfmrod/mapseq
Contact: mering@imls.uzh.ch
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Sequencing the DNA of microbial communities, either wholesale or after amplification of selected marker genes, has greatly advanced our understanding in many fields, including ecology, evolution and medical microbiology. However, the sheer amount of data and the conceptual and technical variability introduced by the wide variety of sequencing and analysis approaches pose difficult challenges for the consolidation and inter-comparability of findings, within and across studies.

The most widely used common denominator for inter-comparisons is taxonomy classification based on ribosomal RNA (rRNA), implemented in a number of software packages including RDP Classifier (Wang *et al.*, 2007), USEARCH (Edgar, 2010), VSEARCH (Rognes *et al.*, 2016) and NINJA-OPS (Al-Ghalith *et al.*, 2016), which are often bundled in broader pipelines such as MOTHUR (Schloss *et al.*, 2009) and QIIME (Caporaso *et al.*, 2010). However, these packages are either restricted to previously known taxa only, are suffering from computational limitations, or cannot be applied in a reference-mapping mode at the scales currently needed. Furthermore, approaches that are restricted to existing taxonomically classified reference sequences may not fully cover microbial diversity. This can be solved by including also unclassified reference sequences, pre-clustered into *operational taxonomic units* (OTUs)—and, ideally, these OTUs should be created at various different identity cutoffs and related to each other hierarchically. Such *hierarchical OTUs* (hOTUs) constitute an operational taxonomy in themselves and enable the assessment of taxa (even uncharacterized taxa) across different studies, at adjustable levels of granularity. MAPseq enables both, by providing a fast and accurate sequence read mapping against hierarchically clustered and annotated reference sequences. In addition to the software itself, we provide a large, curated reference of full-length rRNA genes, pre-clustered into

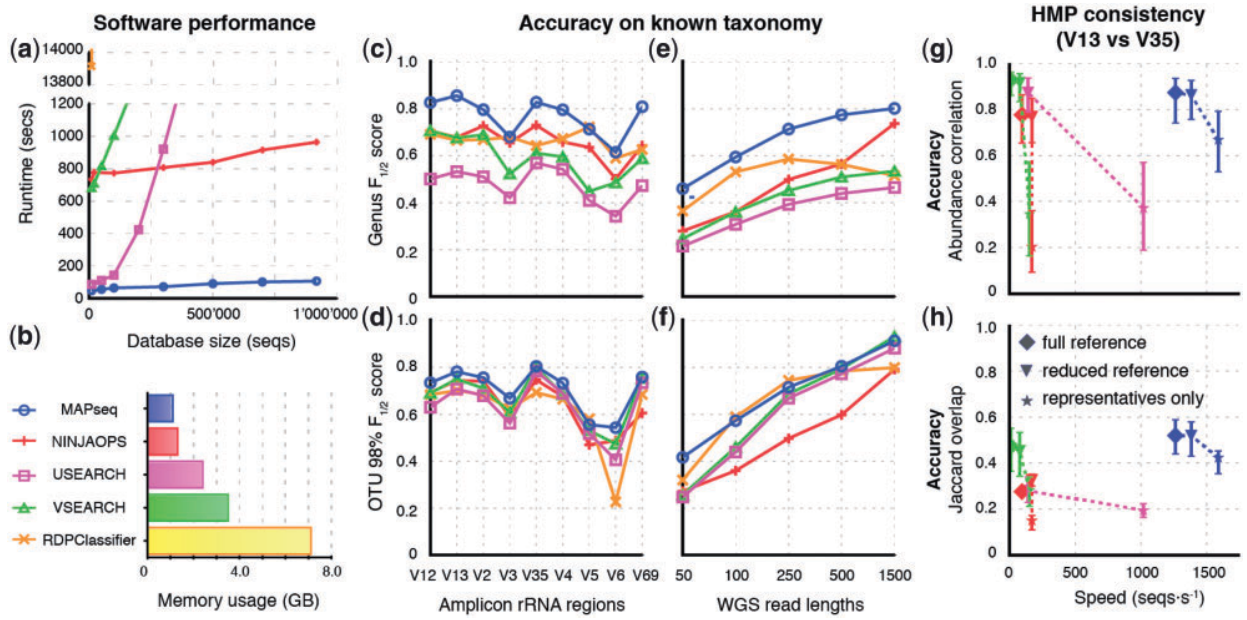


Fig. 1. Benchmarking results on rRNA classification tasks. **(a)** Runtime complexity with increasing database size (using up to 8 CPU threads if supported). **(b)** Maximum memory usage during benchmarking. **(c–f)** $F_{1/2}$-scores for OTU (98%) or taxonomy (genus) classification, for fragments varying in length or rRNA region. Concordance tests on 2, 194 Human Microbiome Project samples, sequenced twice for both the V1V3 and V3V5 regions; **(g)** Pearson correlation of OTU abundances, and **(h)** overlaps in terms of OTUs identified

hOTUs at different identity thresholds, and pre-classified to taxonomic categories based on the NCBI taxonomy and the All-species Living Tree Project dataset (Yilmaz *et al.*, 2014).

2 Results

MAPseq is ten times faster than its closest competitor NINJA-OPS, and a hundred times faster than VSEARCH, on amplicon data (Fig. 1a). Its memory requirements are lower than those of all other tools tested here (Fig. 1b). It can be used also on metagenomic shotgun sequence data, which it automatically searches for suitable rRNA sequences. Accuracy was benchmarked based on placements of reads of known identity, against hOTUs clustered at 98% identity (Fig. 1d and f) as well as against taxonomic categories at genus level (Fig. 1c and e). We also tested MAPseq's performance on reads of different lengths (Fig. 1e and f), as well as different hyper-variable regions of the rRNA gene (Fig. 1c and d).

With few exceptions, MAPseq outperformed all other tools, achieving a maximum $F_{1/2}$-score (weighted harmonic mean of precision and recall) of 0.86 in genus mapping and 0.96 in OTU mapping. The accuracy increase over existing tools is most notable in genus classification, with 30% better $F_{1/2}$-score at 500 bp long reads (Fig. 1e). For all methods, accuracy tends to increase with increasing read lengths, and hyper-variable regions within the rRNA gene yield better placement accuracy than reads originating within random positions within the gene. The increased computational efficiency and higher accuracy of MAPseq are due to several algorithmic innovations, including improved *k-mer* counting based on a pre-clustering step, full Needleman-Wunsch alignment of high scoring segment pairs, and a sensitive algorithm to compute classification confidence; see Supplementary Methods for details.

Strikingly, we find that even small reductions in the comprehensiveness of the reference dataset, by removal of near-identical sequences, can affect the accuracy for all methods tested (Fig. 1g and h; Supplementary Fig. S2). This shows that making reference datasets less redundant for runtime reasons has a significant trade-off cost in terms of mapping accuracy.

For an independent test of classification accuracy, we took advantage of a large data collection for which the very same samples had been subjected to two independent sequencing runs, using different regions of the rRNA gene (2194 samples from the Human Microbiome Project). Here, any good analysis framework should report strongly correlated results. As shown in Figure 1g, the correlations of abundances of mapped OTUs between sequencing runs were found to be fairly high for most methods; however, MAPseq achieved by far the best trade-off in terms of speed versus accuracy. We observed a similar trend in terms of the fraction of shared OTU identifications (Jaccard overlap, Fig. 1h): MAPseq resulted in the highest overlap (median = 0.52), followed by VSEARCH, NINJA-OPS, and USEARCH. Finally, we investigated the effect of using a different method for the *ab initio* clustering of reference sequences into OTUs. Using reference OTUs obtained with UCLUST (a widely adopted method) resulted in lower overall abundance correlations (median = 0.62) and Jaccard overlaps (median = 0.33) (Supplementary Fig. S3), independent of the software subsequently used for the mapping.

As a final validation, we have processed artificial 'mock' community data (Supplementary Fig. S4). We observe that MAPseq recovers their expected abundances better than other tools, at the species, genus and family levels.

In summary, MAPseq outperforms state-of-the-art methods dramatically in terms of speed, while also providing a more accurate and consistent approach. It can be used with the reference data provided, but also with custom references and/or taxonomies. MAPseq is open-source software implemented in multi-threaded C++. Both the software and its reference data are available at: http://www.meringlab.org/software/mapseq/.

Funding

This work was supported by the Swiss National Science Foundation (grant nr. 31003A-160095).

Conflict of Interest: none declared.

References

Al-Ghalith,G.A. *et al*. (2016) NINJA-OPS: fast accurate marker gene alignment using concatenated ribosomes. *PLoS Comput. Biol*., **12**, e1004658.

Caporaso,J.G. *et al*. (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods*, **7**, 335–336.

Edgar,R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460–2461.

Rognes,T. *et al*. (2016) VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, **4**, e2584.

Schloss,P.D. *et al*. (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol*., **75**, 7537–7541.

Wang,Q. *et al*. (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol*., **73**, 5261–5267.

Yilmaz,P. *et al*. (2014) The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. *Nucleic Acids Res*., **42**, D643–D648.

Supplementary Material

Online Methods

Reference dataset

A dataset of publicly available full-length small subunit ribosomal RNA sequences was compiled from the databases RefSeq and NCBI Genbank. Sequences were selected by parsing the Genbank and RefSeq files and extracting any sequences annotated as ribosomal RNA (rRNA keyword) and having 16S or 18S in their annotation information. Sequences smaller than 1000bp, larger than 3500bp, with more than 2% of unknown nucleotides (N) or two such nucleotides in a row (NN) were discarded. Next, sequences were aligned using INFERNAL v1.1.2 with three rRNA models for Bacteria, Archaea and Eukarya, respectively, as available in the ssu-align package (Nawrocki, et al., 2009). The alignment of every sequence against all three models was compared and each sequence was classified as Bacteria, Archaea or Eukarya on the basis of its best alignment score against each of the models. Sequences with a zero or negative alignment score to all models were discarded. Only nucleotides aligned to the models were considered (insertions were removed), and sequences were trimmed between two well-conserved alignment columns found to cover a large fraction of the aligned sequences for each class. Sequences not covering the full length to within 10bp between the trimmed positions were filtered out, as were sequences with more than 30% gaps in the whole aligned region. Chimeric sequences were removed by running UCHIME against a reference set of sequences confirmed by at least three independent studies and falling within the same OTUs (defined at 98%). The final reference set consists of 918'803 sequences. The reference set used for the actual MAPseq mapping process is composed of the unaligned sequences, including the insertions that had been removed during alignment. During the revision of the present manuscript, the reference set was updated (MAPref 2.0) to include 1'585'280 sequences. It had been observerd that many mitochondrial sequences annotated on RefSeq as 16S/18S rRNA sequences did not align to the INFERNAL models used. These are now included in the final reference, nevertheless. Although they could not be assigned OTU labels, they still provide a reference for taxonomic assignments.

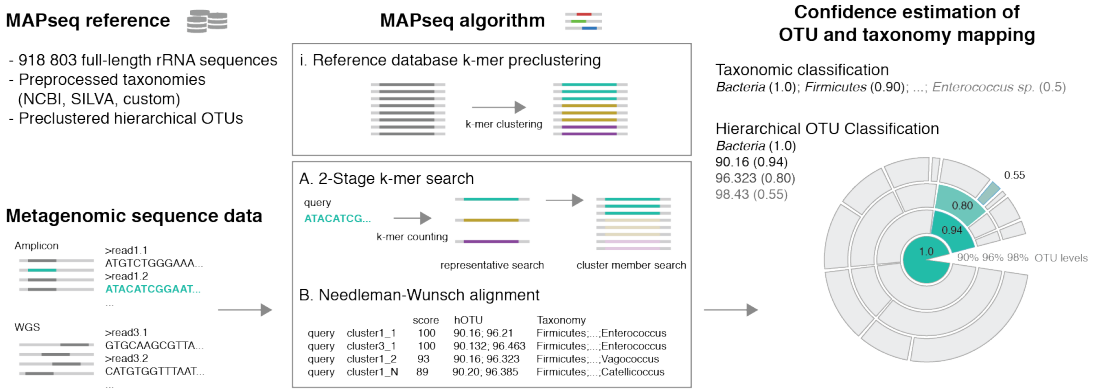


Figure S1 – MAPseq algorithm and reference. Together with the MAPseq software, a large full-length ribosomal RNA reference dataset is provided along with pre-computed hierarchical OTUs and taxonomic assignments from NCBI and SILVA. The algorithm requires a k-mer pre-clustering of the reference dataset which is provided for the MAPseq reference and can be built for custom, user-provided references. The classification of a query sequence begins with a 2-stage k-mer search; first, the representatives of k-mer clusters are searched, and then the members of the top clusters are searched. Finally, the sequences are fully aligned and confidences are estimated for each taxonomic level.

Reference taxonomy

NCBI taxonomies for the reference dataset were assembled either by referring to the annotated taxonomy for sequences extracted from the RefSeq database, or by scanning for GenBank entries annotated to be culture collection strains. This taxonomy reference set comprised 89,315 sequences, covering 15,810 species, 3,637 genera, 1,493 families, 800 orders, 404 classes, and 185 phyla. Another, independent taxonomy annotation was obtained from SILVA Living Tree Project (LTP) (Yilmaz, et al., 2014) by mapping the sequences annotated in the SILVA LTP database to the MAPseq reference set. The set of sequences with a trusted taxonomy was defined as the *gold* set and was used to predict the taxonomy of the remaining, un-annotated sequences. The updated reference (MAPref 2.0) covers in the NCBI taxonomy: 16’323 species, 6’981 genera, 2’654 families, 1’091 orders, 412 classes, and 175 phyla.

Hierarchical OTUs

The set of aligned sequences in the Archaea, Bacteria and Eukarya reference datasets were separately clustered using an average-linkage hierarchical clustering algorithm implemented in HPC-CLUST (Matias Rodrigues and von Mering, 2014) down to 90% sequence identity. Each sequence was then assigned to an OTU at five different identity cutoffs: 99%, 98%, 96% and 90%, yielding 144’596, 92’862, 51’904, and 19’957 OTUs, respectively. The updated reference (MAPref 2.0) includes one additionally level at the 97% identity cutoff, for a total of 6 levels.

MAPseq algorithm

MAPseq achieves a big advance in efficiency and accuracy at searching databases of highly similar sequences, by improving upon the k-mer counting approach used in many other sequence search tools such as BLAST or USEARCH. To achieve this efficiency, the reference dataset is pre-clustered into clusters of sequences. This pre-clustering is made available to

MAPseq together with the actual sequence data and the taxonomic information. At runtime, MAPseq uses a 2-stage k-mer search – first to the k-mer cluster representatives, and second to the cluster members – to effectively reduce the number of sequences to be searched exponentially and to significantly improve the memory requirements of the program.

K-mer pre-clustering

Sequences in the datasets to be used with MAPseq are pre-clustered on the basis of the number of shared k-mers, in two phases. The first sequence in the dataset is used as the seed for the first cluster, afterwards, every sequence in the dataset is iteratively compared on the basis of shared k-mers to the existing cluster representatives and are either assigned to one of the existing clusters or used as the representative for a new cluster. Sequences are added to existing clusters when the sequence being considered shares at least 80% of its k-mers with a cluster representative. The first phase of pre-clustering is complete once all sequences are assigned to a cluster or used as representatives for new clusters. In the second phase, all cluster representatives are kept from the first phase, but the membership assignment for the other sequences is recalculated, this ensures that sequences assigned to a cluster at some stage can be assigned to another cluster representative to which they are more similar. This situation can occur because cluster representatives can be created at any point after sequences have already been earlier assigned to other clusters in the first phase.
Pre-clustering of our dataset consisting of 918’803 sequences results in 56’169 k-mer clusters which implies a 16 times reduction in number of sequences that need to be searched in the first stage of the k-mer search step. In the updated reference (MAPref 2.0), the clustering of 1’585’280 sequences yielded 91’181 clusters.

Mapping of sequence reads to the reference dataset

When a query is first searched against the database, the number of shared k-mers between the query and one representative for each cluster is computed. In a second step, the number of shared k-mers is computed between the query and all of the members belonging to the clusters that ranked highest in the previous step. Finally, the top hits are aligned to the query sequence. Dynamic programming is used to identify the set of identical segments yielding the largest alignment score, followed by alignment of the regions between these segments using the banded Needleman-Wunsch algorithm. The ends of the alignment are determined using the banded Smith-Waterman algorithm with drop-off such that the alignment is not extended if the alignment score drops to more than 20 below best score found to that point. The best match is taken as the best guess for taxonomy or OTU assignment, and the remaining hits are used to estimate the confidence in the assignment at each level.

Estimating assignment confidence

MAPseq achieves higher accuracy by assigning a mapping confidence to better control false classifications. There are two different types of false classifications: i) misclassifications due to missing sequences in the reference dataset, and ii) incorrect assignments among the existing sequences in the database. These two types of errors are controlled for independently. The first case is controlled by computing a confidence on the basis of identity cutoffs previously

optimized for each level for a given taxonomy and reference dataset. The identity confidence is computed using the following formula:

$$cf_{id}(i) = \frac{(id_i - id_{cutoff} + 0.02)}{\Delta_{cutoff}}.$$

Where id_i is the identity of the query compared to the reference sequence *i*, id_{cutoff} is the identity cutoff, and Δ_{cutoff} is a parameter that controls the strength of the effect of identity differences on the confidence; these two parameters were optimized previously for each taxonomic level. The additional 0.02 was added so that the id_{cutoff} term matches the OTU cutoffs in OTU mapping.
The second case is controlled using the formula described below that weighs the score of the top hit against subsequent hits to different taxa:

$$cf_{score}(i) = \frac{w_i}{W}, \qquad w_i = e^{\alpha(1-\frac{T}{S_i})}, \qquad W = \sum_i w_i.$$

Where S_i is the alignment score of reference sequence *i* in the top hit list, *T* is the score of the best aligned sequence, w_i is the weight of sequence *i*, *W* is the sum over all weights, α is a parameter controlling the influence of lower scoring hits. In essence, the higher the difference in scores between the top hit and the closest second-best hits to any conflicting taxonomy/OTU, the better the confidence placed in the assignment. This automatically solves the problem of sequence reads originating in highly conserved parts of the 16S rRNA sequence, because top hits will tend to have similar scores even when belonging to conflicting taxonomies. The final confidence is calculated as the minimum between the two confidences:

$$cf(i) = min(cf_{score}(i), cf_{id}(i)).$$

Another advantage of this approach is that higher confidences are automatically computed for lower taxonomic levels, since second-best hits will necessarily have higher score differences at lower taxonomic levels.

Validation of sequence read mapping

Benchmarking of the taxonomic classification and OTU mapping performance was done using as a starting point a set of nearly full-length 16S/18S ribosomal subunit sequences compiled from the Genbank database. The taxonomy from the All-Species Living Tree Project (LTP) (Yarza, et al., 2010) which comprises a set of manually curated sequence and taxonomies was used as the gold standard in our benchmarks. For the OTU mapping, sequences were clustered using HPC-CLUST (Matias Rodrigues and von Mering, 2014) after alignment with the INFERNAL aligner (Nawrocki, et al., 2009), or with UCLUST.

To benchmark the accuracy of MAPseq and other classification tools commonly used in metagenomic data analysis (RDP Classifier 2.6, NINJA-OPS 1.5, VSEARCH 2.4.0 and USEARCH 9.2.64) we generated a benchmark query and reference dataset from our full-length rRNA gene dataset in which the true OTU assignments or taxonomies were known. We avoided benchmarking trivial cases when the query and database sequence were nearly identical, by excluding from the benchmark reference any sequences originating from the same species or

OTU (at 99% identity cutoff) as any of the query sequences. In addition, we evaluated the ability of different methods to correctly ignore sequences that had no representative in the reference database by generating the reference set such that in the best case only 50% of the query sequences actually had a representative of their genus or OTU (at 98% identity cutoff). This test set was generated by randomly selecting one sequence for each genus from the full set while respecting the previous conditions.

Confidence thresholds per method were chosen as the thresholds yielding maximum $F_{1/2}$-scores averaged over the scores obtained for different read lengths and different rRNA hypervariable regions. For USEARCH and VSEARCH, which output no confidences, we used the difference above the identity cutoff for each level as a measure of confidence, for example when a query had 98.1% identity to a reference sequence then it had 0.1 confidence when mapping to an OTU at 98% identity cutoff.

The query sets were generated by segmenting the selected sequences into different lengths (50bp, 100bp, 250bp, 500bp, and 1500bp) and at different positions incrementally in steps of half the segment length. For example, 50bp segments were generated starting at positions (1, 25, 50, 75, …) and 100bp were generated starting at positions (1, 50, 100, …). For the hypervariable regions, the sequences were trimmed at positions matching the *Escherichia coli* 16S rRNA after alignment with INFERNAL. The hypervariable regions were selected according to (Schloss, 2010) as follows: V2 (100bp to 337bp), V3 (357bp to 514bp), V4 (578bp to 784bp), V5 (784bp to 986bp), V6 (986bp to 1045bp), V1V2 (28bp to 337bp), V1V3 (28bp to 514bp), V3V5 (357bp to 906bp) and V6V9 (986bp to 1491bp).

The same test reference dataset and queries were used with all methods tested.

Computation speed and memory benchmarks

The computation speed and memory benchmarks were performed by running each tool with 8 threads (except RDP classifier which does not support multithreading) on the same dedicated Dell Blade M605 computer with 2 quad-core Opteron 2.33 GHz processors and 24 GB of random access memory. For the benchmark (Figure 1a), the input data used was the Human Microbiome Project sequencing run (700016012) consisting of amplicon raw data targeting the V3V5 region of the 16S rRNA gene. This sample was mapped to different subsamples of the MAPseq reference database consisting of 918 803 sequences, and in (Figure 1g,h) the input data used was a 10% subsampled set of all raw reads found in 2,194 HMP samples for which both V1-V3 and V3-V5 sequencing runs existed.
USEARCH could not be run on reference datasets larger than 650’000 sequences, due to memory usage limitations of the 32bit version (the only freely available version). The following commands were used for running the benchmarks:

USEARCH:

```
usearch9.2.64 -usearch_local input.fa -db mapseqref.udb9.2 -threads 8 -id 0.90 -strand
both -blast6out usearch.output
```

VSEARCH:

```
vsearch -usearch_global input.fa -db mapseqref.fa -threads 8 -id 0.90 -strand
both -blast6out vsearch.output
```

NINJA-OPS:

```
python NINJA-OPS/bin/ninja.py -d 0 -p 8 -z -i input.fa -b mapseqref
```

MAPseq:

```
mapseq -nthreads 8 input.fa mapseqref.fasta
```

RDP Classifier:

```
java -Xmx10g -jar rdp_classifier_2.6/dist/classifier.jar classify -t
mapseqref.rdpdb/rRNAClassifier.properties -o rdpclass.output input.fa
```

Concordance analysis of independent V1V3 and V3V5 sequence data of the Human Microbiome Project

Raw 16S rRNA V1V3 and V3V5 amplicon sequencing data and sample metadata of the Human Microbiome Project (HMP) were downloaded from the NCBI Sequence Read Archive and the HMP data depository (hmpdacc.org). Chimeric reads were detected using UCHIME (Edgar, et al., 2011) in both *de novo* mode and with a custom in-house reference database of non-chimeric sequences; reads labeled as chimeric by both approaches were removed from further analyses. Filtered reads were then aligned to a 16S rRNA model using INFERNAL (Nawrocki, et al., 2009) and pruned to the respective alignment flanking positions for the V1V3 and V3V5 primer sets, as described above. Reads that did not align to these regions, that had too many gaps within flanking positions or that were not observed in at least two samples independently at a 5bp error tolerance were removed from further analyses. Moreover, all biological samples were removed that did not contain at least 1,000 sequences of both V1V3 and V3V5. After these filtering steps, there remained 2,194 samples containing a total of 17,890,946 (V1V3) and 26,627,383 (V3V5) sequences.

Sequences were assigned to OTUs by reference mapping. Consistency between V1V3 and V3V5 data from the same biological samples was assessed as direct per-sample Jaccard similarity (i.e., the overlap of reference OTUs at 97% called in both V1V3 and V3V5 sequencing of the same sample) and as the Pearson correlation of abundances of OTUs shared between both sequencing sets per sample. Reference mapping was performed on three different reference databases: One including all full-length sequences for each OTU (full reference, Figure 1g,h), one with 30% randomly picked representatives per OTU (reduced reference) and one including a single random representative for each OTU (representatives only). Speed estimates are based on reduced input files (random 10% of the sequences for each sample) and tool parameters as in the previous section, while consistency (Jaccard similarity, Pearson correlation) was computed on all input sequences with 40 to 80 threads on a larger workstation.

Supplementary Results

Accuracy reduction when using representative reference datasets

A common practice used in ribosomal RNA marker gene analysis is to make the reference database non-redundant or use only representatives of each OTU. In Fig. S2 we show that this practice leads up to 0.1 less F-score than using the full reference set (Fig. 1f). This reduction in $F_{1/2}$-score was not limited to MAPseq but was also observed for all other tools tested.

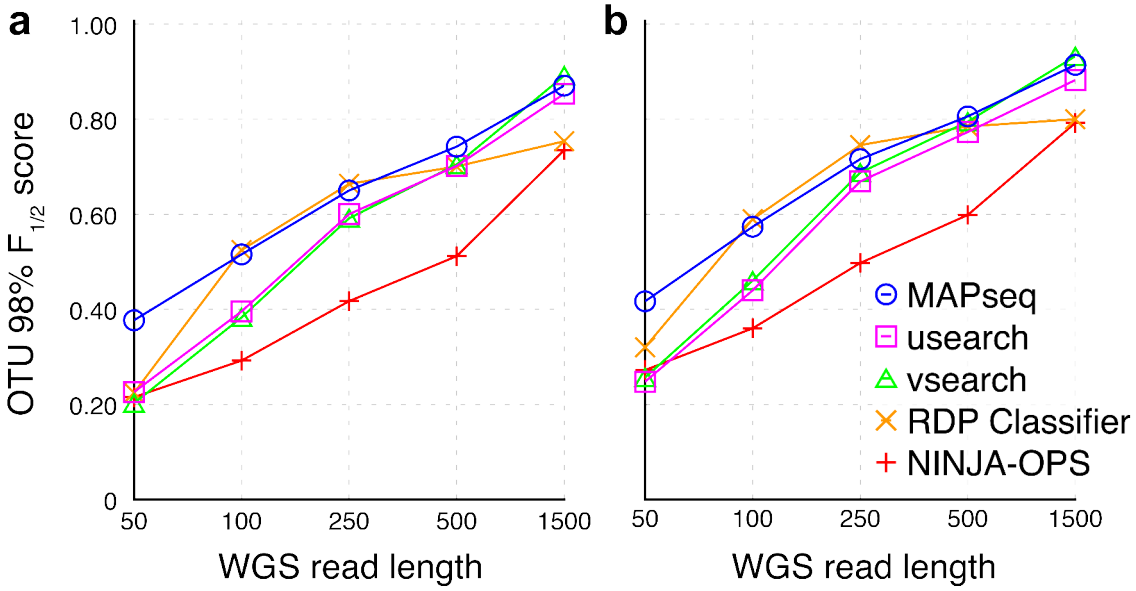


Figure S2 – $F_{1/2}$ scores for sequence reads of different lengths, mapped to OTUs at 98% cutoff using: a) non-redundant/representative reference dataset and b) the full reference, identical to Fig. 1f and included here for direct comparison.

HPC-CLUST based OTUs outperform UCLUST OTUs in mapping consistency

We chose a subset of HMP metagenomic data in which the same samples had been sequenced twice, targeting two different regions of the rRNA gene (V1 to V3 and V3 to V5); these were in total 2,194 samples. Abundance correlations for mapped OTUs at 97% identity between sequenced subregions per sample were very high (median=0.83, mean=0.73; Figure S2a) when using MAPseq and our reference OTU clustering. When using MAPseq to map to the same reference (but pre-clustered by UCLUST, the tool used in the GreenGenes and SILVA databases) the correlation was lower (median=0.62, mean=0.59), and when using UCLUST (v1.2.22q) to map the sequences the correlations were yet significantly lower both when mapping to our reference (median=0.39, mean=0.41) as well as to the UCLUST-clustered reference (median=0.22, mean=0.29; this latter approach corresponds to the default for “closed-reference OTU picking” in the widely used QIIME pipeline). We observed a similar trend in the fraction of identified OTUs common to both pairs of sequencing runs for the same sample (Figure S3b): MAPseq mapping to the MAPseq reference resulted in the highest overlap (median=0.46, mean=0.43), followed by MAPseq mapping to the UCLUST-clustered reference

(median=0.33, mean=0.31). Using UCLUST resulted in a much-reduced overlap between OTUs common to pairs of sequencing runs when mapping to OTUs (median=0.30, mean=0.29) and when mapping to UCLUST representatives (median=0.23, mean=0.22).

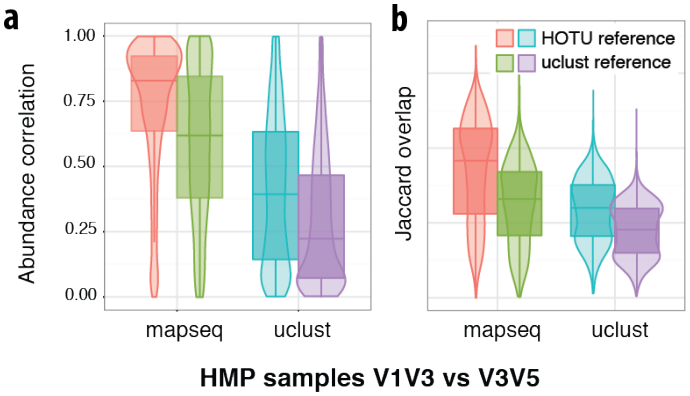


Figure S3 – Abundance correlations and Jaccard overlaps for OTUs identified in pairs of sequencing runs of identical HMP samples targeting two different ribosomal RNA regions (V1-V3 and V3-V5). The same reference dataset was used with two different clustering approaches: hierarchical clustering using average-linkage computed with HPC-CLUST, and a UCLUST clustering.

Finally, we compared these two OTU clustering methods to *de novo* clustering of the metagenomic samples. OTU set compositional consistency between reference-mapped partitions and their respective *de novo* counterparts was checked by calculating the Adjusted Mutual Information (AMI, (Vinh, et al., 2009)) between the respective OTU sets (e.g., comparing MAPseq against an average linkage pre-clustered reference with an average linkage *de novo* clustered partition). Ideally, the results of *de novo* clustering and reference-mapping strategies should correspond to each other – if consistent sequence processing and clustering algorithms are applied. However, this correspondence has recently been called into question for UCLUST (Westcott and Schloss, 2015). To test this, we quantified partition similarity in terms of OTU composition (per-sequence cluster membership) as *Adjusted Mutual Information* (AMI); AMI values of 1 indicate perfect partition identity, and AMI values of zero indicate random compositional agreement as expected by chance. For the HMP dataset, we observed that MAPseq against an average linkage (AL) OTU reference provided a good correspondence with AL *de novo* clustering (AMI=0.83). This indicates that MAPseq reference mapping, although computationally much more efficient, may indeed approximate hierarchical AL *de novo* clustering, which is arguably still a gold standard in marker gene processing. In contrast, UCLUST mapping against a UCLUST-clustered non-redundant reference (the default method in QIIME, see above) showed much lower agreement with *de novo* UCLUST clustering (AMI=0.66). UCLUST performed better against our AL OTU reference (AMI=0.75), but was outperformed by MAPseq against a UCLUST reference (AMI=0.79). Thus, both reference pre-clustering and choice of mapping tool have a strong effect on consistency, and UCLUST was clearly outperformed on both, by MAPseq mapping and by AL clustering.

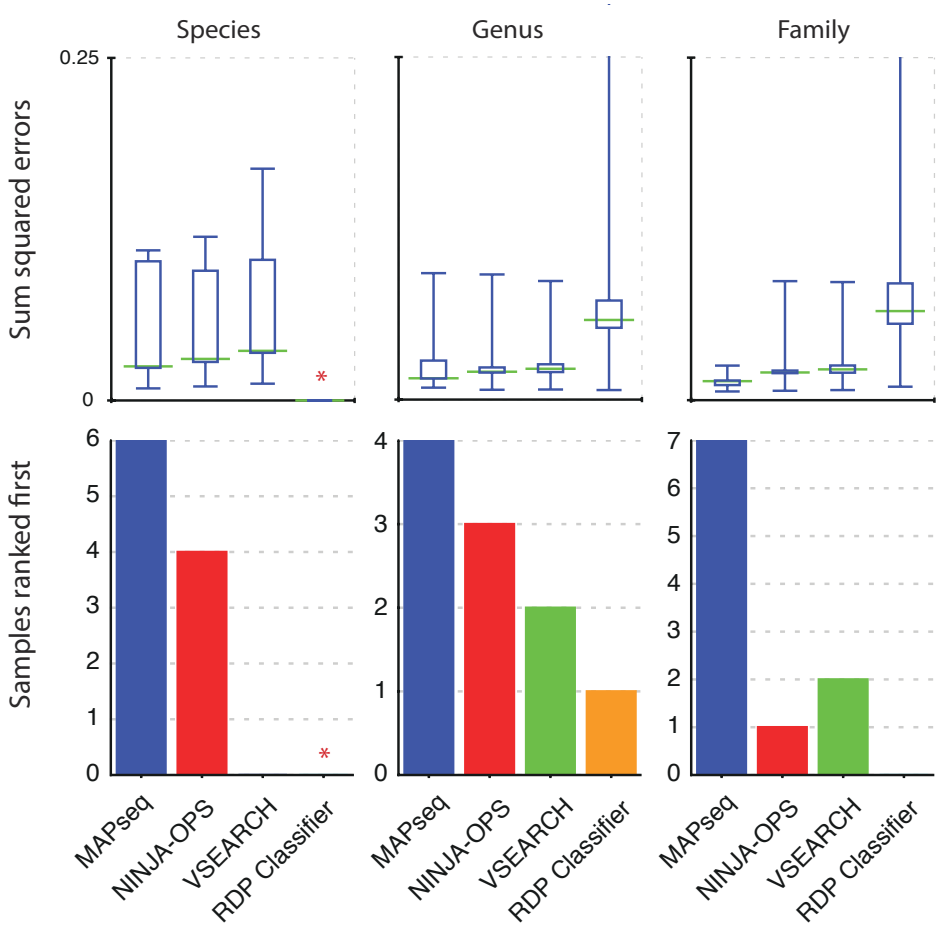


Figure S4 – Analysis of ten mock communities. Top panels, box plots of sum of squared errors between predicted and expected taxa abundance. Bottom panels, number of times each tool ranked first over the ten mock communities. *) RDP classifier does not report species classifications and therefore no result could be computed at the species level.

MAPseq outperforms NINJA-OPS, VSEARCH, and RDP Classifier in the analysis of mock communities.

We downloaded ten mock communities available in the Mockrobiota dataset (Bokulich, et al., 2016), specifically: (Gohl, et al., 2016; Kozich, et al., 2013; Tourlousse, et al., 2017). After downloading, we mapped the forward reads to our full reference using MAPseq, NINJA-OPS, VSEARCH, and RDP Classifier. As a measure of performance, we computed the sum of squared errors (SSE) between predicted and expected abundances. Fig. S4 shows the box plots of the SSE obtained for each tool over all the sequence runs and how often each tool ranked first over all samples for the species, genus, and family taxonomic levels. MAPseq outperformed the other tools, obtaining an overall smaller median of the SSE and consistently ranking first more frequently than other tools.

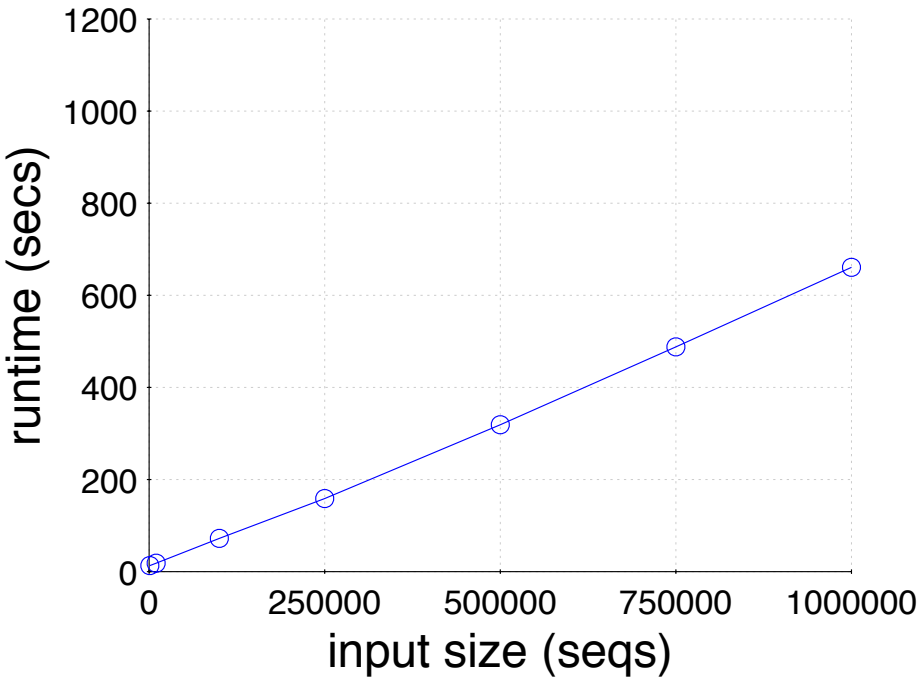


Figure S5 – Analysis of MAPseq runtime complexity with increasing input size.

MAPseq exhibits linear complexity with input size.

We benchmarked the time it took to analyse several downsamplings of the HMP dataset used in the concordance analysis of the V13 and V35 sequencing runs. The downsampled queries ranged between a thousand reads and a million reads. Fig. S5 shows that the MAPseq runtime is practically linear when using 8 cores over a large range of input sizes. The full MAPseq reference was used as the mapping target.

References

Bokulich, N.A., *et al.* mockrobiota: a Public Resource for Microbiome Bioinformatics Benchmarking. *mSystems* 2016;1(5).

Edgar, R.C., *et al.* UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* 2011;27(16):2194-2200.

Gohl, D.M., *et al.* Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies. *Nat Biotechnol* 2016;34(9):942-949.

Kozich, J.J., *et al.* Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Applied and environmental microbiology* 2013;79(17):5112-5120.

Matias Rodrigues, J.F. and von Mering, C. HPC-CLUST: distributed hierarchical clustering for large sets of nucleotide sequences. *Bioinformatics* 2014;30(2):287-288.

Nawrocki, E.P., Kolbe, D.L. and Eddy, S.R. Infernal 1.0: inference of RNA alignments. *Bioinformatics* 2009;25(10):1335-1337.

Schloss, P.D. The effects of alignment quality, distance calculation method, sequence filtering, and region on the analysis of 16S rRNA gene-based studies. *PLoS computational biology* 2010;6(7):e1000844.

Tourlousse, D.M.*, et al.* Synthetic spike-in standards for high-throughput 16S rRNA gene amplicon sequencing. *Nucleic acids research* 2017;45(4):e23.

Vinh, N.X., Epps, J. and Bailey, J. Information theoretic measure for clusterings comparison: is correction for chance necessary? In, *26th Annual International Conference on Machine Learning*. ACM, Montreal, Canada; 2009. p. pp. 10.1145/1553374.1553511.

Westcott, S.L. and Schloss, P.D. De novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units. *PeerJ* 2015;3:e1487.

Yarza, P.*, et al.* Update of the All-Species Living Tree Project based on 16S and 23S rRNA sequence analyses. *Systematic and applied microbiology* 2010;33(6):291-299.

Yilmaz, P.*, et al.* The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. *Nucleic acids research* 2014;42(Database issue):D643-648.

Appendix B

Curriculum Vitae

Curriculum Vitae

Janko Tackmann

Born August 13th 1987 in Kyritz, Germany

Professional Experience

November 2014 –present	**PhD student at the Life Science Zurich Graduate School (Switzerland)** — *Institute of Molecular Life Sciences* at University Zürich, group *of Prof. Dr. Christian von Mering*
February 2014 –May 2014	**Research internship at the Biological Research Centre, Szeged (Hungary)** — Department *Synthetic and Systems Biology*, group of *Dr. Balász Papp*
March 2013 –July 2013	**Master's project at the Biomedical Centre, Uppsala (Sweden)** — Department *Molecular Evolution*, group of *Dr. Thijs Ettema*
August 2011 –Octobre 2011	**Research internship at the Max-Planck-Institute for Evolutionary Anthropology, Leipzig (Germany)** — Department *Evolutionary Genetics*, group of *Dr. Daniel Falush*
April 2011 –June 2011	**Practical at the DFG Research Center MATHEON, Berlin (Germany)** — Group *Mathematics in Life Sciences*, lead by *Prof. Dr. Alexander Bockmayr*

Education

2014 –present	**PhD in Bioinformatics and Microbial Ecology, Universität Zürich (Switzerland)**
2011–2014	**MSc in Bioinformatics, Freie Universität Berlin (Germany)** — *including one year of studies in Evolutionary Biology at Uppsala Universitetet (Sweden)*
2008–2011	**BSc in Bioinformatics, Freie Universität Berlin (Germany)**
2007	**Abitur (the German A level), majoring in Computer Science and Mathematics**

Bibliography

Abadi, Martín, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. (2016). "Tensorflow: a system for large-scale machine learning." In: *OSDI*. Vol. 16, pp. 265–283.

Abbeele, Pieter Van den, Tom Van de Wiele, Willy Verstraete, and Sam Possemiers (2011). "The host selects mucosal and luminal associations of coevolved gut microorganisms: a novel concept." In: *FEMS Microbiology Reviews* 35.4, pp. 681–704.

Achtman, Mark and Michael Wagner (2008). "Microbial diversity and the genetic nature of microbial species." In: *Nature Reviews. Microbiology* 6.6, pp. 431–440.

Ackermann, Martin (2015). "A functional perspective on phenotypic heterogeneity in microorganisms." In: *Nature Reviews Microbiology* 13.8, pp. 497–508.

Adams, James B., Leah J. Johansen, Linda D. Powell, David Quig, and Robert A. Rubin (2011). "Gastrointestinal flora and gastrointestinal status in children with autism–comparisons to typical children and correlation with autism severity." In: *BMC gastroenterology* 11, p. 22.

Agler, Matthew T., Jonas Ruhe, Samuel Kroll, Constanze Morhenn, Sang-Tae Kim, Detlef Weigel, and Eric M. Kemen (2016). "Microbial Hub Taxa Link Host and Abiotic Factors to Plant Microbiome Variation." In: *PLoS biology* 14.1, e1002352.

Aitchison, John (1981). "A new approach to null correlations of proportions." In: *Journal of the International Association for Mathematical Geology* 13.2, pp. 175–189.

Aitchison, John (1982). "The statistical analysis of compositional data." In: *Journal of the Royal Statistical Society: Series B (Methodological)* 44.2, pp. 139–160.

Aitchison, John, Carles Barceló-Vidal, José Antonio Martín-Fernández, and Vera Pawlowsky-Glahn (2000). "Logratio analysis and compositional distance." In: *Mathematical Geology* 32.3, pp. 271–275.

Albert, null, null Jeong, and null Barabasi (2000). "Error and attack tolerance of complex networks." In: *Nature* 406.6794, pp. 378–382.

Albert, Reka and Albert-Laszlo Barabasi (2002). "Statistical mechanics of complex networks." In: *Reviews of Modern Physics* 74.1, pp. 47–97.

Aliferis, Constantin F, Alexander Statnikov, Ioannis Tsamardinos, Subramani Mani, and Xenofon D Koutsoukos (2010a). "Local causal and markov blanket induction

for causal discovery and feature selection for classification part i: Algorithms and empirical evaluation." In: *Journal of Machine Learning Research* 11.Jan, pp. 171–234.

Aliferis, Constantin F, Alexander Statnikov, Ioannis Tsamardinos, Subramani Mani, and Xenofon D Koutsoukos (2010b). "Local causal and markov blanket induction for causal discovery and feature selection for classification part ii: Analysis and extensions." In: *Journal of Machine Learning Research* 11.Jan, pp. 235–284.

Aliferis, Constantin F, Ioannis Tsamardinos, and Alexander Statnikov (2003). "HITON: a novel Markov Blanket algorithm for optimal variable selection." In: *AMIA Annual Symposium Proceedings*. Vol. 2003. American Medical Informatics Association, p. 21.

Allison, Steven D. and Jennifer B. H. Martiny (2008). "Resistance, resilience, and redundancy in microbial communities." In: *Proceedings of the National Academy of Sciences* 105.Supplement 1, pp. 11512–11519.

Alvarez-Acosta, Thais, Cira León, Salvador Acosta-González, Haydeé Parra-Soto, Isabel Cluet-Rodriguez, Maria Rosario Rossell, and José A. Colina-Chourio (2009). "Beneficial role of green plantain [Musa paradisiaca] in the management of persistent diarrhea: a prospective randomized trial." In: *Journal of the American College of Nutrition* 28.2, pp. 169–176.

Amaral, Luıs A Nunes, Antonio Scala, Marc Barthelemy, and H Eugene Stanley (2000). "Classes of small-world networks." In: *Proceedings of the national academy of sciences* 97.21, pp. 11149–11152.

Amir, Amnon, Daniel McDonald, Jose A Navas-Molina, Evguenia Kopylova, James T Morton, Zhenjiang Zech Xu, Eric P Kightley, Luke R Thompson, Embriette R Hyde, Antonio Gonzalez, et al. (2017). "Deblur rapidly resolves single-nucleotide community sequence patterns." In: *MSystems* 2.2, e00191–16.

Anderson, Marti J. (2001). "A new method for non-parametric multivariate analysis of variance." In: *Austral Ecology* 26.1, pp. 32–46.

Andrieu, Christophe, Nando De Freitas, Arnaud Doucet, and Michael I Jordan (2003). "An introduction to MCMC for machine learning." In: *Machine learning* 50.1-2, pp. 5–43.

Angulo, Marco Tulio, Jaime A Moreno, Gabor Lippner, Albert-László Barabási, and Yang-Yu Liu (2017). "Fundamental limitations of network reconstruction from temporal data." In: *Journal of the Royal Society Interface* 14.127, p. 20160966.

Arumugam, Manimozhiyan, Jeroen Raes, Eric Pelletier, Denis Le Paslier, Takuji Yamada, Daniel R Mende, Gabriel R Fernandes, Julien Tap, Thomas Bruls, Jean-Michel Batto, et al. (2011). "Enterotypes of the human gut microbiome." In: *nature* 473.7346, p. 174.

Askew, RR (1961). "On the biology of the inhabitants of oak galls of Cynipidae (Hymenoptera) in Britain." In: *Trans Soc Bri Entomol* 14, pp. 237–268.

Bäckhed, Fredrik (2012). "Host responses to the human microbiome." In: *Nutrition reviews* 70.suppl_1, S14–S17.

Baldwin, Bruce G, Michael J Sanderson, J Mark Porter, Martin F Wojciechowski, Christopher S Campbell, and Michael J Donoghue (1995). "The its Region of Nuclear Ribosomal DNA: A Valuable Source of Evidence on Angiosperm Phylogeny." In: *Ann. Mo. Bot. Gard.* 82.2, p. 247.

Balvočiūtė, Monika and Daniel H Huson (2017). "SILVA, RDP, Greengenes, NCBI and OTT—how do these taxonomies compare?" In: *BMC genomics* 18.2, p. 114.

Banerjee, Samiran, Klaus Schlaeppi, and Marcel GA Heijden (2018). "Keystone taxa as drivers of microbiome structure and functioning." In: *Nature Reviews Microbiology*, p. 1.

Bar-On, Yinon M, Rob Phillips, and Ron Milo (2018). "The biomass distribution on Earth." In: *Proceedings of the National Academy of Sciences* 115.25, pp. 6506–6511.

Barabás, György, Matthew J Michalska-Smith, and Stefano Allesina (2017). "Self-regulation and the stability of large ecological networks." In: *Nature ecology & evolution* 1.12, p. 1870.

Barabási, Albert-László and Réka Albert (1999). "Emergence of scaling in random networks." In: *science* 286.5439, pp. 509–512.

Barberán, Albert, Scott T Bates, Emilio O Casamayor, and Noah Fierer (2012). "Using network analysis to explore co-occurrence patterns in soil microbial communities." In: *ISME J.* 6.2, pp. 343–351.

Bartlett, John MS and David Stirling (2003). "A short history of the polymerase chain reaction." In: *PCR protocols*. Springer, pp. 3–6.

Bascompte, Jordi (2009). "Disentangling the web of life." In: *Science* 325.5939, pp. 416–419.

Bascompte, Jordi and Pedro Jordano (2007). "Plant-Animal Mutualistic Networks: The Architecture of Biodiversity." In: *Annual Review of Ecology, Evolution, and Systematics* 38.1, pp. 567–593.

Bascompte, Jordi, Pedro Jordano, Carlos J. Melián, and Jens M. Olesen (2003). "The nested assembly of plant–animal mutualistic networks." In: *Proceedings of the National Academy of Sciences* 100.16, pp. 9383–9387.

Bascompte, Jordi, Pedro Jordano, and Jens M Olesen (2006). "Asymmetric coevolutionary networks facilitate biodiversity maintenance." In: *Science* 312.5772, pp. 431–433.

Bashiardes, Stavros, Gili Zilberman-Schapira, and Eran Elinav (2016). "Use of metatranscriptomics in microbiome research." In: *Bioinformatics and biology insights* 10, BBI–S34610.

Bastolla, Ugo, Miguel A Fortuna, Alberto Pascual-García, Antonio Ferrera, Bartolo Luque, and Jordi Bascompte (2009). "The architecture of mutualistic networks minimizes competition and increases biodiversity." In: *Nature* 458.7241, p. 1018.

Bauer, Maria A, Katharina Kainz, Didac Carmona-Gutierrez, and Frank Madeo (2018). "Microbial wars: Competition in ecological niches and within the microbiome." In: *Microbial Cell* 5.5, p. 215.

Beaumont, Michelle, Julia K. Goodrich, Matthew A. Jackson, Idil Yet, Emily R. Davenport, Sara Vieira-Silva, Justine Debelius, Tess Pallister, Massimo Mangino, Jeroen Raes, Rob Knight, Andrew G. Clark, Ruth E. Ley, Tim D. Spector, and Jordana T. Bell (2016). "Heritable components of the human fecal microbiome are associated with visceral fat." In: *Genome Biology* 17.1, p. 189.

Belenguer, Alvaro, Sylvia H Duncan, A Graham Calder, Grietje Holtrop, Petra Louis, Gerald E Lobley, and Harry J Flint (2006). "Two routes of metabolic cross-feeding between Bifidobacterium adolescentis and butyrate-producing anaerobes from the human gut." In: *Appl. Environ. Microbiol.* 72.5, pp. 3593–3599.

Bellman, Richard (1961). *Adaptive control processes: A guided tour*.

Benotti, Mark J. and Bruce J. Brownawell (2009). "Microbial degradation of pharmaceuticals in estuarine and coastal seawater." In: *Environmental Pollution* 157.3, pp. 994–1002.

Benson, Andrew K, Scott A Kelly, Ryan Legge, Fangrui Ma, Soo Jen Low, Jaehyoung Kim, Min Zhang, Phaik Lyn Oh, Derrick Nehrenberg, Kunjie Hua, et al. (2010). "Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors." In: *Proceedings of the National Academy of Sciences* 107.44, pp. 18933–18938.

Berry, David and Stefanie Widder (2014). "Deciphering microbial interactions and detecting keystone species with co-occurrence networks." In: *Frontiers in microbiology* 5, p. 219.

Berry, Michelle A, Jeffrey D White, Timothy W Davis, Sunit Jain, Thomas H Johengen, Gregory J Dick, Orlando Sarnelle, and Vincent J Denef (2017). "Are oligotypes meaningful ecological and phylogenetic units? A case study of Microcystis in freshwater lakes." In: *Frontiers in microbiology* 8, p. 365.

Bertrand, Jean-Claude, Pierre Caumette, Philippe Lebaron, Robert Matheron, Philippe Normand, and Télesphore Sime-Ngando (2015). *Environmental microbiology: fundamentals and applications*. Springer.

Besard, Tim, Christophe Foket, and Bjorn De Sutter (2018). "Effective extensible programming: Unleashing julia on gpus." In: *IEEE Transactions on Parallel and Distributed Systems*.

Bezanson, Jeff, Alan Edelman, Stefan Karpinski, and Viral B Shah (2017). "Julia: A fresh approach to numerical computing." In: *SIAM review* 59.1, pp. 65–98.

Bianchi, Micheline, Danielle Marty, Jean-Louis Teyssié, and Scott W Fowler (1992). "Strictly aerobic and anaerobic bacteria associated with sinking particulate matter and zooplankton fecal pellets." In: *Marine Ecology Progress Series*, pp. 55–60.

Bikel, Shirley, Alejandra Valdez-Lara, Fernanda Cornejo-Granados, Karina Rico, Samuel Canizales-Quinteros, Xavier Soberón, Luis Del Pozo-Yauner, and Adrián Ochoa-Leyva (2015). "Combining metagenomics, metatranscriptomics and viromics to explore novel microbial interactions: towards a systems-level understanding of human microbiome." In: *Computational and Structural Biotechnology Journal* 13, pp. 390–401.

Blei, David M, Alp Kucukelbir, and Jon D McAuliffe (2017). "Variational Inference: A Review for Statisticians." In: *J. Am. Stat. Assoc.* 112.518, pp. 859–877.

Bleidorn, Christoph (2016). "Third generation sequencing: technology and its potential impact on evolutionary biodiversity research." In: *Systematics and Biodiversity* 14.1, pp. 1–8.

Bodenhausen, Natacha, Miriam Bortfeld-Miller, Martin Ackermann, and Julia A. Vorholt (2014). "A synthetic community approach reveals plant genotypes affecting the phyllosphere microbiota." In: *PLoS genetics* 10.4, e1004283.

Bohan, David A, Alan Raybould, Christian Mulder, Guy Woodward, Alireza Tamaddoni-Nezhad, Nico Bluthgen, Michael JO Pocock, Stephen Muggleton, Darren M Evans, Julia Astegiano, et al. (2013). "Networking agroecology: integrating the diversity of agroecosystem interactions." In: *Advances in Ecological Research*. Vol. 49. Elsevier, pp. 1–67.

Bonnet, Régis, Antonia Suau, Joël Doré, Glenn R. Gibson, and Matthew D. Collins (2002). "Differences in rDNA libraries of faecal bacteria derived from 10- and 25-cycle PCRs." In: *International Journal of Systematic and Evolutionary Microbiology* 52.Pt 3, pp. 757–763.

Bouguelia, Sihem, Yoann Roupioz, Sami Slimani, Laure Mondani, Maria G. Casabona, Claire Durmort, Thierry Vernet, Roberto Calemczuk, and Thierry Livache (2013). "On-chip microbial culture for the specific detection of very low levels of bacteria." In: *Lab on a Chip* 13.20, pp. 4024–4032.

Bray, J Roger and John T Curtis (1957). "An ordination of the upland forest communities of southern Wisconsin." In: *Ecological monographs* 27.4, pp. 325–349.

Breiman, Leo (2001). "Random Forests." In: *Mach. Learn.* 1.45, pp. 5–32.

Brierley, Corale L. and James A. Brierley (1973). "A chemoautotrophic and thermophilic microorganism isolated from an acid hot spring." In: *Canadian Journal of Microbiology* 19.2, pp. 183–188.

Brooks, Andrew W., Kevin D. Kohl, Robert M. Brucker, Edward J. van Opstal, and Seth R. Bordenstein (2016). "Phylosymbiosis: Relationships and Functional Effects of

Microbial Communities across Host Evolutionary History." In: *PLOS Biology* 14.11, e2000225.

Brown, Christopher T., Laura A. Hug, Brian C. Thomas, Itai Sharon, Cindy J. Castelle, Andrea Singh, Michael J. Wilkins, Kelly C. Wrighton, Kenneth H. Williams, and Jillian F. Banfield (2015). "Unusual biology across a group comprising more than 15% of domain Bacteria." In: *Nature* 523.7559, pp. 208–211.

Bruns, Alke, Herbert Hoffelner, and Jörg Overmann (2003). "A novel approach for high throughput cultivation assays and the isolation of planktonic bacteria." In: *FEMS Microbiology Ecology* 45.2, pp. 161–171.

Bucci, Vanni, Belinda Tzen, Ning Li, Matt Simmons, Takeshi Tanoue, Elijah Bogart, Luxue Deng, Vladimir Yeliseyev, Mary L Delaney, Qing Liu, Bernat Olle, Richard R Stein, Kenya Honda, Lynn Bry, and Georg K Gerber (2016). "MDSINE: Microbial Dynamical Systems INference Engine for microbiome time-series analyses." In: *Genome Biol.* 17.1, p. 121.

Buffie, Charlie G., Vanni Bucci, Richard R. Stein, Peter T. McKenney, Lilan Ling, Asia Gobourne, Daniel No, Hui Liu, Melissa Kinnebrew, Agnes Viale, Eric Littmann, Marcel R. M. van den Brink, Robert R. Jenq, Ying Taur, Chris Sander, Justin R. Cross, Nora C. Toussaint, Joao B. Xavier, and Eric G. Pamer (2015). "Precision microbiome reconstitution restores bile acid mediated resistance to Clostridium difficile." In: *Nature* 517.7533, pp. 205–208.

Buffie, Charlie G. and Eric G. Pamer (2013). "Microbiota-mediated colonization resistance against intestinal pathogens." In: *Nature reviews. Immunology* 13.11, pp. 790–801.

Bühlmann, Peter, Markus Kalisch, and Lukas Meier (2014). "High-Dimensional Statistics with a View Toward Applications in Biology." In: *Annual Review of Statistics and Its Application* 1.1, pp. 255–278.

Buntine, Wray (1996). "A guide to the literature on learning probabilistic networks from data." In: *IEEE Transactions on knowledge and data engineering* 8.2, pp. 195–210.

Byrne, CS, ES Chambers, DJ Morrison, and G Frost (2015). "The role of short chain fatty acids in appetite regulation and energy homeostasis." In: *International journal of obesity* 39.9, p. 1331.

Callahan, Benjamin J, Paul J McMurdie, and Susan P Holmes (2017). "Exact sequence variants should replace operational taxonomic units in marker-gene data analysis." In: *ISME J.* 11.12, pp. 2639–2643.

Callahan, Benjamin J, Paul J McMurdie, Michael J Rosen, Andrew W Han, Amy Jo A Johnson, and Susan P Holmes (2016). "DADA2: High-resolution sample inference from Illumina amplicon data." In: *Nat. Methods* 13.7, pp. 581–583.

Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L Madden (2009). "BLAST+: architecture and applications." In: *BMC bioinformatics* 10.1, p. 421.

Camarinha-Silva, Amélia, Ruy Jáuregui, Diego Chaves-Moreno, Andrew P. A. Oxley, Frieder Schaumburg, Karsten Becker, Melissa L. Wos-Oxley, and Dietmar H. Pieper (2014). "Comparing the anterior nare bacterial community of two discrete human populations using Illumina amplicon sequencing." In: *Environmental Microbiology* 16.9, pp. 2939–2952.

Cani, Patrice D and Willem M de Vos (2017). "Next-generation beneficial microbes: the case of Akkermansia muciniphila." In: *Frontiers in microbiology* 8, p. 1765.

Cao, Hong-Tai, Travis E Gibson, Amir Bashan, and Yang-Yu Liu (2017). "Inferring human microbial dynamics from temporal metagenomics data: Pitfalls and lessons." In: *BioEssays* 39.2, p. 1600188.

Cao, Yang, Yuanyuan Wang, Xiaofei Zheng, Fei Li, and Xiaochen Bo (2016). "RevEcoR: an R package for the reverse ecology analysis of microbiomes." In: *BMC bioinformatics* 17.1, p. 294.

Caporaso, J Gregory, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D Bushman, Elizabeth K Costello, Noah Fierer, Antonio Gonzalez Peña, Julia K Goodrich, Jeffrey I Gordon, Gavin A Huttley, Scott T Kelley, Dan Knights, Jeremy E Koenig, Ruth E Ley, Catherine A Lozupone, Daniel McDonald, Brian D Muegge, Meg Pirrung, Jens Reeder, Joel R Sevinsky, Peter J Turnbaugh, William A Walters, Jeremy Widmann, Tanya Yatsunenko, Jesse Zaneveld, and Rob Knight (2010). "QIIME allows analysis of high-throughput community sequencing data." In: *Nat. Methods* 7.5, pp. 335–336.

Caporaso, J Gregory, Christian L Lauber, Elizabeth K Costello, Donna Berg-Lyons, Antonio Gonzalez, Jesse Stombaugh, Dan Knights, Pawel Gajer, Jacques Ravel, Noah Fierer, et al. (2011). "Moving pictures of the human microbiome." In: *Genome biology* 12.5, R50.

Carabotti, Marilia, Annunziata Scirocco, Maria Antonietta Maselli, and Carola Severi (2015). "The gut-brain axis: interactions between enteric microbiota, central and enteric nervous systems." In: *Annals of gastroenterology: quarterly publication of the Hellenic Society of Gastroenterology* 28.2, p. 203.

Carbonero, Franck, Ann C Benefiel, and H Rex Gaskins (2012). "Contributions of the microbial hydrogen economy to colonic homeostasis." In: *Nat. Rev. Gastroenterol. Hepatol.* 9.9, pp. 504–518.

Cardona, Silvia, Anat Eck, Montserrat Cassellas, Milagros Gallart, Carmen Alastrue, Joel Dore, Fernando Azpiroz, Joaquim Roca, Francisco Guarner, and Chaysavanh Manichanh (2012). "Storage conditions of intestinal microbiota matter in metagenomic analysis." In: *BMC Microbiology* 12, p. 158.

Caspi, Ron, Kate Dreher, and Peter D Karp (2013). "The challenge of constructing, classifying, and representing metabolic pathways." In: *FEMS Microbiol. Lett.* 345.2, pp. 85–93.

Chaffron, Samuel, Hubert Rehrauer, Jakob Pernthaler, and Christian von Mering (2010). "A global network of coexisting microbes from environmental and whole-genome sequence data." In: *Genome Research* 20.7, pp. 947–959.

Chao, Anne (1984). "Nonparametric Estimation of the Number of Classes in a Population." In: *Scandinavian Journal of Statistics* 11.4, pp. 265–270.

Chao, Anne, Robin L Chazdon, Robert K Colwell, and Tsung-Jen Shen (2004). "A new statistical approach for assessing similarity of species composition with incidence and abundance data." In: *Ecol. Lett.* 8.2, pp. 148–159.

Chao, Anne, Chun-Huo Chiu, and Lou Jost (2014). "Unifying Species Diversity, Phylogenetic Diversity, Functional Diversity, and Related Similarity and Differentiation Measures Through Hill Numbers." In: *Annu. Rev. Ecol. Evol. Syst.* 45.1, pp. 297–324.

Chao, Anne and Lou Jost (2012). "Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size." In: *Ecology* 93.12, pp. 2533–2547.

Chapman, Harold W. (1997). "Comparative Physiology of the Vertebrate Digestive System, 2nd ed." In: *The Canadian Veterinary Journal* 38.9, pp. 576–577.

Chen, Eric Z and Hongzhe Li (2016). "A two-part mixed-effects model for analyzing longitudinal microbiome compositional data." In: *Bioinformatics* 32.17, pp. 2611–2617.

Chen, I-Min A, Victor M Markowitz, Ken Chu, Krishna Palaniappan, Ernest Szeto, Manoj Pillay, Anna Ratner, Jinghua Huang, Evan Andersen, Marcel Huntemann, Neha Varghese, Michalis Hadjithomas, Kristin Tennessen, Torben Nielsen, Natalia N Ivanova, and Nikos C Kyrpides (2017). "IMG/M: integrated genome and metagenome comparative data analysis system." In: *Nucleic Acids Res.* 45.D1, pp. D507–D516.

Chickering, David Maxwell (2002). "Optimal structure identification with greedy search." In: *Journal of machine learning research* 3.Nov, pp. 507–554.

Chickering, David Maxwell, David Heckerman, and Christopher Meek (2004). "Large-sample learning of Bayesian networks is NP-hard." In: *Journal of Machine Learning Research* 5.Oct, pp. 1287–1330.

Chodkowski, John L and Ashley Shade (2017). "A synthetic community system for probing microbial interactions driven by exometabolites." In: *MSystems* 2.6, e00129–17.

Chuang, Han-Yu, Matan Hofree, and Trey Ideker (2010). "A Decade of Systems Biology." In: *Annu. Rev. Cell Dev. Biol.* 26.1, pp. 721–744.

Claesson, Marcus J, Siobhán Cusack, Orla O'Sullivan, Rachel Greene-Diniz, Heleen de Weerd, Edel Flannery, Julian R Marchesi, Daniel Falush, Timothy Dinan, Gerald Fitzgerald, et al. (2011). "Composition, variability, and temporal stability of the intestinal microbiota of the elderly." In: *Proceedings of the National Academy of Sciences* 108.Supplement 1, pp. 4586–4591.

Clark, Nicholas J, Konstans Wells, and Oscar Lindberg (2018). "Unravelling changing interspecific interactions across environmental gradients using Markov random fields." In: *Ecology* 99.6, pp. 1277–1283.

Clauset, Aaron, Cristopher Moore, and Mark EJ Newman (2008). "Hierarchical structure and the prediction of missing links in networks." In: *Nature* 453.7191, p. 98.

Clavel, Thomas, Joël Doré, and Michael Blaut (2006). "Bioavailability of lignans in human subjects." In: *Nutrition Research Reviews* 19.2, pp. 187–196.

Coelho, Luis Pedro, Renato Alves, Paulo Monteiro, Jaime Huerta-Cepas, Ana Teresa Freitas, and Peer Bork (2018). "NG-meta-profiler: fast processing of metagenomes using NGLess, a domain-specific language." In: *BioRxiv*, p. 367755.

Cole, James R, Qiong Wang, Jordan A Fish, Benli Chai, Donna M McGarrell, Yanni Sun, C Titus Brown, Andrea Porras-Alfaro, Cheryl R Kuske, and James M Tiedje (2014). "Ribosomal Database Project: data and tools for high throughput rRNA analysis." In: *Nucleic Acids Res.* 42.Database issue, pp. D633–42.

Cordier, Tristan, Dominik Forster, Yoann Dufresne, Catarina I M Martins, Thorsten Stoeck, and Jan Pawlowski (2018). "Supervised machine learning outperforms taxonomy-based environmental DNA metabarcoding applied to biomonitoring." In: *Mol. Ecol. Resour.* 18.6, pp. 1381–1391.

Costea, Paul I, Georg Zeller, Shinichi Sunagawa, and Peer Bork (2014). "A fair comparison." In: *Nat. Methods* 11.4, p. 359.

Costea, Paul Igor, Robin Munch, Luis Pedro Coelho, Lucas Paoli, Shinichi Sunagawa, and Peer Bork (2017). "metaSNV: a tool for metagenomic strain level analysis." In: *PLoS One* 12.7, e0182392.

Cox, Cymon J, Peter G Foster, Robert P Hirt, Simon R Harris, and T Martin Embley (2008). "The archaebacterial origin of eukaryotes." In: *Proceedings of the National Academy of Sciences of the United States of America* 105.51, pp. 20356–61.

Coyte, Katharine Z., Jonas Schluter, and Kevin R. Foster (2015). "The ecology of the microbiome: Networks, competition, and stability." In: *Science* 350.6261, pp. 663–666.

Cregger, MA, AM Veach, ZK Yang, MJ Crouch, R Vilgalys, GA Tuskan, and CW Schadt (2018). "The Populus holobiont: dissecting the effects of plant niches and genotype on the microbiome." In: *Microbiome* 6.1, p. 31.

Cuevas, Daniel A, Janaka Edirisinghe, Chris S Henry, Ross Overbeek, Taylor G O' Connell, and Robert A Edwards (2016). "From DNA to FBA: How to Build Your Own Genome-Scale Metabolic Model." In: *Front. Microbiol.* 7, p. 907.

Curtis, Meredith M., Zeping Hu, Claire Klimko, Sanjeev Narayanan, Ralph Deberardinis, and Vanessa Sperandio (2014). "The gut commensal Bacteroides thetaiotaomicron exacerbates enteric infection through modification of the metabolic landscape." In: *Cell Host & Microbe* 16.6, pp. 759–769.

Dai, Zhenwei, Sunny H Wong, Jun Yu, and Yingying Wei (2018). "Batch effects correction for microbiome data with Dirichlet-multinomial regression." In: *Bioinformatics*.

Dam, Phuongan, Luis L Fonseca, Konstantinos T Konstantinidis, and Eberhard O Voit (2016). "Dynamic models of the complex microbial metapopulation of lake mendota." In: *NPJ Syst Biol Appl* 2, p. 16007.

Darwin, Charles (1859). *On the origin of species by means of natural selection, or preservation of favoured races in the struggle for life*. London :John Murray.

Das, Nilanjana and Preethy Chandran (2011). "Microbial degradation of petroleum hydrocarbon contaminants: an overview." In: *Biotechnology research international* 2011.

David, Lawrence A, Arne C Materna, Jonathan Friedman, Maria I Campos-Baptista, Matthew C Blackburn, Allison Perrotta, Susan E Erdman, and Eric J Alm (2014a). "Host lifestyle affects human microbiota on daily timescales." In: *Genome Biol.* 15.7, R89.

David, Lawrence A., Corinne F. Maurice, Rachel N. Carmody, David B. Gootenberg, Julie E. Button, Benjamin E. Wolfe, Alisha V. Ling, A. Sloan Devlin, Yug Varma, Michael A. Fischbach, Sudha B. Biddinger, Rachel J. Dutton, and Peter J. Turnbaugh (2014b). "Diet rapidly and reproducibly alters the human gut microbiome." In: *Nature* 505.7484, pp. 559–563.

Deacon, Jim W (2013). *Fungal biology*. John Wiley & Sons.

Dean, Frank B., Seiyu Hosono, Linhua Fang, Xiaohong Wu, A. Fawad Faruqi, Patricia Bray-Ward, Zhenyu Sun, Qiuling Zong, Yuefen Du, Jing Du, Mark Driscoll, Wanmin Song, Stephen F. Kingsmore, Michael Egholm, and Roger S. Lasken (2002). "Comprehensive human genome amplification using multiple displacement amplification." In: *Proceedings of the National Academy of Sciences of the United States of America* 99.8, pp. 5261–5266.

Debelius, Justine, Se Jin Song, Yoshiki Vazquez-Baeza, Zhenjiang Zech Xu, Antonio Gonzalez, and Rob Knight (2016). "Tiny microbes, enormous impacts: what matters in gut microbiome studies?" In: *Genome biology* 17.1, p. 217.

Dedysh, Svetlana N (2011). "Cultivating uncultured bacteria from northern wetlands: knowledge gained and remaining gaps." In: *Frontiers in microbiology* 2, p. 184.

DeGruttola, Arianna K, Daren Low, Atsushi Mizoguchi, and Emiko Mizoguchi (2016). "Current Understanding of Dysbiosis in Disease in Human and Animal Models." In: *Inflamm. Bowel Dis.* 22.5, pp. 1137–1150.

DeJongh, Matthew, Kevin Formsma, Paul Boillot, John Gould, Matthew Rycenga, and Aaron Best (2007). "Toward the automated generation of genome-scale metabolic networks in the SEED." In: *BMC Bioinformatics* 8, p. 139.

Delmas, Eva, Mathilde Besson, Marie-Hélène Brice, Laura A Burkle, Giulio V Dalla Riva, Marie-Josée Fortin, Dominique Gravel, Paulo R Guimarães Jr, David H Hembry, Erica A Newman, et al. (2019). "Analysing ecological networks of species interactions." In: *Biological Reviews* 94.1, pp. 16–36.

Deng, Ye, Yi-Huei Jiang, Yunfeng Yang, Zhili He, Feng Luo, and Jizhong Zhou (2012). "Molecular ecological network analyses." In: *BMC Bioinformatics* 13.1, p. 113.

Desai, Mahesh S., Anna M. Seekatz, Nicole M. Koropatkin, Nobuhiko Kamada, Christina A. Hickey, Mathis Wolter, Nicholas A. Pudlo, Sho Kitamoto, Nicolas Terrapon, Arnaud Muller, Vincent B. Young, Bernard Henrissat, Paul Wilmes, Thaddeus S. Stappenbeck, Gabriel Núñez, and Eric C. Martens (2016). "A Dietary Fiber-Deprived Gut Microbiota Degrades the Colonic Mucus Barrier and Enhances Pathogen Susceptibility." In: *Cell* 167.5, 1339–1353.e21.

DeSantis, Todd Z, Philip Hugenholtz, Neils Larsen, Mark Rojas, Eoin L Brodie, Keith Keller, Thomas Huber, Daniel Dalevi, Ping Hu, and Gary L Andersen (2006). "Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB." In: *Appl. Environ. Microbiol.* 72.7, pp. 5069–5072.

Desbrosses, Guilhem J. and Jens Stougaard (2011). "Root Nodulation: A Paradigm for How Plant-Microbe Symbiosis Influences Host Developmental Pathways." In: *Cell Host & Microbe* 10.4, pp. 348–358.

Dethlefsen, Les, Sue Huse, Mitchell L. Sogin, and David A. Relman (2008). "The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing." In: *PLoS biology* 6.11, e280.

Dethlefsen, Les and David A. Relman (2011). "Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation." In: *Proceedings of the National Academy of Sciences of the United States of America* 108 Suppl 1, pp. 4554–4561.

Diamond, Jared M (1975). "Assembly of species communities." In: *Ecology and evolution of communities*, pp. 342–444.

Diels, Ludo and Max Mergeay (1990). "DNA Probe-Mediated Detection of Resistant Bacteria from Soils Highly Polluted by Heavy Metals." In: *Appl. Environ. Microbiol.* 56.5, pp. 1485–1491.

Diwan, Vaibhav, Hans-Jørgen Albrechtsen, Barth F. Smets, and Arnaud Dechesne (2018). "Does universal 16S rRNA gene amplicon sequencing of environmental communities

provide an accurate description of nitrifying guilds?" In: *Journal of Microbiological Methods* 151, pp. 28–34.

Dixon, Ray and Daniel Kahn (2004). "Genetic regulation of biological nitrogen fixation." In: *Nature Reviews Microbiology* 2.8, pp. 621–631.

Donati, Claudio, Moreno Zolfo, Davide Albanese, Duy Tin Truong, Francesco Asnicar, Valerio Iebba, Duccio Cavalieri, Olivier Jousson, Carlotta De Filippo, Curtis Huttenhower, et al. (2016). "Uncovering oral Neisseria tropism and persistence using metagenomic sequencing." In: *Nature microbiology* 1.7, p. 16070.

Doney, Scott C., Mary Ruckelshaus, J. Emmett Duffy, James P. Barry, Francis Chan, Chad A. English, Heather M. Galindo, Jacqueline M. Grebmeier, Anne B. Hollowed, Nancy Knowlton, Jeffrey Polovina, Nancy N. Rabalais, William J. Sydeman, and Lynne D. Talley (2012). "Climate Change Impacts on Marine Ecosystems." In: *Annual Review of Marine Science* 4.1, pp. 11–37.

Doolittle, W Ford and Eric Bapteste (2007). "Pattern pluralism and the Tree of Life hypothesis." In: *Proceedings of the National Academy of Sciences of the United States of America* 104.7, pp. 2043–9.

Doolittle, W. Ford and R. Thane Papke (2006). "Genomics and the bacterial species problem." In: *Genome Biology* 7.9, p. 116.

Doolittle, W. Ford and Olga Zhaxybayeva (2009). "On the origin of prokaryotic species." In: *Genome Research* 19.5, pp. 744–756.

D' Onofrio, Anthony, Jason M. Crawford, Eric J. Stewart, Kathrin Witt, Ekaterina Gavrish, Slava Epstein, Jon Clardy, and Kim Lewis (2010). "Siderophores from Neighboring Organisms Promote the Growth of Uncultured Bacteria." In: *Chemistry & Biology* 17.3, pp. 254–264.

Dubilier, Nicole, Claudia Bergin, and Christian Lott (2008). "Symbiotic diversity in marine animals: the art of harnessing chemosynthesis." In: *Nature Reviews Microbiology* 6.10, pp. 725–740.

Duda, Stephany, Constantin Aliferis, Randolph Miller, Alexander Statnikov, and Kevin Johnson (2005). "Extracting drug-drug interaction articles from MEDLINE to improve the content of drug databases." In: *AMIA Annu. Symp. Proc.*, pp. 216–220.

Edgar, Robert C (2010). "Search and clustering orders of magnitude faster than BLAST." In: *Bioinformatics* 26.19, pp. 2460–2461.

Edgar, Robert C (2013). "UPARSE: highly accurate OTU sequences from microbial amplicon reads." In: *Nat. Methods* 10.10, pp. 996–998.

Edgar, Robert C and Henrik Flyvbjerg (2015). "Error filtering, pair assembly and error correction for next-generation sequencing reads." In: *Bioinformatics* 31.21, pp. 3476–3482.

Egozcue, Juan José (2018). “Linear Association in Compositional Data Analysis.” In: *Austrian Journal of Statistics* 47.1, p. 3.

Egozcue, Juan José, Vera Pawlowsky-Glahn, Glòria Mateu-Figueras, and Carles Barceló-Vidal (2003). “Isometric logratio transformations for compositional data analysis.” In: *Mathematical Geology* 35.3, pp. 279–300.

Eichorst, Stephanie A., Florian Strasser, Tanja Woyke, Arno Schintlmeister, Michael Wagner, and Dagmar Woebken (2015). “Advancements in the application of NanoSIMS and Raman microspectroscopy to investigate the activity of microbial cells in soils.” In: *FEMS Microbiology Ecology* 91.10.

Erdos, Paul and Alfréd Rényi (1960). “On the evolution of random graphs.” In: *Publ. Math. Inst. Hung. Acad. Sci* 5.1, pp. 17–60.

Eren, A Murat, Hilary G Morrison, Pamela J Lescault, Julie Reveillaud, Joseph H Vineis, and Mitchell L Sogin (2015). “Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences.” In: *ISME J.* 9.4, pp. 968–979.

Ereshefsky, Marc (2010). “Microbiology and the species problem.” In: *Biol. Philos.* 25.4, pp. 553–568.

Faith, Daniel P (1992). “Conservation evaluation and phylogenetic diversity.” In: *Biol. Conserv.* 61.1, pp. 1–10.

Faith, Jeremiah J., Federico E. Rey, David O' Donnell, Maria Karlsson, Nathan P. McNulty, George Kallstrom, Andrew L. Goodman, and Jeffrey I. Gordon (2010). “Creating and characterizing communities of human gut microbes in gnotobiotic mice.” In: *The ISME journal* 4.9, pp. 1094–1098.

Falkowski, Paul G., Tom Fenchel, and Edward F. Delong (2008). “The Microbial Engines That Drive Earth's Biogeochemical Cycles.” In: *Science* 320.5879, pp. 1034–1039.

Falony, Gwen, Marie Joossens, Sara Vieira-Silva, Jun Wang, Youssef Darzi, Karoline Faust, Alexander Kurilshikov, Marc Jan Bonder, Mireia Valles-Colomer, Doris Vandeputte, Raul Y Tito, Samuel Chaffron, Leen Rymenans, Chloë Verspecht, Lise De Sutter, Gipsi Lima-Mendez, Kevin D' hoe, Karl Jonckheere, Daniel Homola, Roberto Garcia, Ettje F Tigchelaar, Linda Eeckhaudt, Jingyuan Fu, Liesbet Henckaerts, Alexandra Zhernakova, Cisca Wijmenga, and Jeroen Raes (2016). “Population-level analysis of gut microbiome variation.” In: *Science* 352.6285, pp. 560–564.

Falony, Gwen, Angeliki Vlachou, Kristof Verbrugghe, and Luc De Vuyst (2006). “Cross-feeding between Bifidobacterium longum BB536 and acetate-converting, butyrate-producing colon bacteria during growth on oligofructose.” In: *Applied and Environmental Microbiology* 72.12, pp. 7835–7841.

Fang, Huaying, Chengcheng Huang, Hongyu Zhao, and Minghua Deng (2015). “CCLasso: correlation inference for compositional data through Lasso.” In: *Bioinformatics* 31.19, pp. 3172–3180.

Fang, Huaying, Chengcheng Huang, Hongyu Zhao, and Minghua Deng (2017). “gCoda: Conditional Dependence Network Inference for Compositional Data.” In: *J. Comput. Biol.* 24.7, pp. 699–708.

Farquhar, James, Huiming Bao, and Mark Thiemens (2000). “Atmospheric Influence of Earth's Earliest Sulfur Cycle.” In: *Science* 289.5480, pp. 756–758.

Farris, MH and JB Olson (2007). “Detection of Actinobacteria cultivated from environmental samples reveals bias in universal primers.” In: *Letters in applied microbiology* 45.4, pp. 376–381.

Faust, Karoline, Franziska Bauchinger, Béatrice Laroche, Sophie de Buyl, Leo Lahti, Alex D Washburne, Didier Gonze, and Stefanie Widder (2018). “Signatures of ecological processes in microbial community time series.” In: *Microbiome* 6.1, p. 120.

Faust, Karoline, Leo Lahti, Didier Gonze, Willem M de Vos, and Jeroen Raes (2015). “Metagenomics meets time series analysis: unraveling microbial community dynamics.” In: *Curr. Opin. Microbiol.* 25, pp. 56–66.

Faust, Karoline and Jeroen Raes (2012). “Microbial interactions: from networks to models.” In: *Nature Reviews Microbiology* 10.8, p. 538.

Faust, Karoline and Jeroen Raes (2016). “CoNet app: inference of biological association networks using Cytoscape.” In: *F1000Res.* 5, p. 1519.

Faust, Karoline, J Fah Sathirapongsasuti, Jacques Izard, Nicola Segata, Dirk Gevers, Jeroen Raes, and Curtis Huttenhower (2012). “Microbial co-occurrence relationships in the human microbiome.” In: *PLoS computational biology* 8.7, e1002606.

Federhen, S (2011). “The NCBI Taxonomy database.” In: *Nucleic Acids Res.* 40.D1, pp. D136–D143.

Feichtmayer, Judith, Li Deng, and Christian Griebler (2017). “Antagonistic microbial interactions: contributions and potential applications for controlling pathogens in the aquatic systems.” In: *Frontiers in microbiology* 8, p. 2192.

Feigelman, Rounak, Christian R Kahlert, Florent Baty, Frank Rassouli, Rebekka L Kleiner, Philipp Kohler, Martin H Brutsche, and Christian von Mering (2017). “Sputum DNA sequencing in cystic fibrosis: non-invasive access to the lung microbiome and to pathogen details.” In: *Microbiome* 5.1, p. 20.

Fernandes, Andrew D, Jennifer Ns Reid, Jean M Macklaim, Thomas A McMurrough, David R Edgell, and Gregory B Gloor (2014). “Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis.” In: *Microbiome* 2, p. 15.

Field, Christopher B., Michael J. Behrenfeld, James T. Randerson, and Paul Falkowski (1998). “Primary Production of the Biosphere: Integrating Terrestrial and Oceanic Components.” In: *Science* 281.5374, pp. 237–240.

Fierer, Noah, Christian L. Lauber, Nick Zhou, Daniel McDonald, Elizabeth K. Costello, and Rob Knight (2010). "Forensic identification using skin bacterial communities." In: *Proceedings of the National Academy of Sciences* 107.14, pp. 6477–6481.

Filippo, Carlotta De, Duccio Cavalieri, Monica Di Paola, Matteo Ramazzotti, Jean Baptiste Poullet, Sebastien Massart, Silvia Collini, Giuseppe Pieraccini, and Paolo Lionetti (2010). "Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa." In: *Proceedings of the National Academy of Sciences* 107.33, pp. 14691–14696.

Fischbach, Michael A. and Justin L. Sonnenburg (2011). "Eating For Two: How Metabolism Establishes Interspecies Interactions in the Gut." In: *Cell Host & Microbe* 10.4, pp. 336–347.

Fischer, Keno and Elliot Saba (2018). "Automatic Full Compilation of Julia Programs and ML Models to Cloud TPUs." In: *arXiv:1810.09868 [cs, stat]*.

Fisher, Charles K and Pankaj Mehta (2014). "Identifying keystone species in the human gut microbiome from metagenomic timeseries using sparse linear regression." In: *PLoS One* 9.7, e102451.

Flint, Harry J., Edward A. Bayer, Marco T. Rincon, Raphael Lamed, and Bryan A. White (2008). "Polysaccharide utilization by gut bacteria: potential for new insights from genomic analysis." In: *Nature Reviews Microbiology* 6.2, pp. 121–131.

Flombaum, Pedro, José L. Gallegos, Rodolfo A. Gordillo, José Rincón, Lina L. Zabala, Nianzhi Jiao, David M. Karl, William K. W. Li, Michael W. Lomas, Daniele Veneziano, Carolina S. Vera, Jasper A. Vrugt, and Adam C. Martiny (2013). "Present and future global distributions of the marine Cyanobacteria Prochlorococcus and Synechococcus." In: *Proceedings of the National Academy of Sciences of the United States of America* 110.24, pp. 9824–9829.

Foesel, Bärbel U., Verena Nägele, Astrid Naether, Pia K. Wüst, Jan Weinert, Michael Bonkowski, Gertrud Lohaus, Andrea Polle, Fabian Alt, Yvonne Oelmann, Markus Fischer, Michael W. Friedrich, and Jörg Overmann (2014). "Determinants of Acidobacteria activity inferred from the relative abundances of 16S rRNA transcripts in German grassland and forest soils." In: *Environmental Microbiology* 16.3, pp. 658–675.

Forslund, Kristoffer, Falk Hildebrand, Trine Nielsen, Gwen Falony, Emmanuelle Le Chatelier, Shinichi Sunagawa, Edi Prifti, Sara Vieira-Silva, Valborg Gudmundsdottir, Helle K Pedersen, Manimozhiyan Arumugam, Karsten Kristiansen, Anita Yvonne Voigt, Henrik Vestergaard, Rajna Hercog, Paul Igor Costea, Jens Roat Kultima, Junhua Li, Torben Jørgensen, Florence Levenez, Joël Dore, MetaHIT consortium, H Bjørn Nielsen, Søren Brunak, Jeroen Raes, Torben Hansen, Jun Wang, S Dusko Ehrlich, Peer Bork, and Oluf Pedersen (2015). "Disentangling type 2 diabetes and

metformin treatment signatures in the human gut microbiota." In: *Nature* 528.7581, pp. 262–266.

Forster, Dominik, Sabine Filker, Rebecca Kochems, Hans-Werner Breiner, Tristan Cordier, Jan Pawlowski, and Thorsten Stoeck (2018). "A Comparison of Different Ciliate Metabarcode Genes as Bioindicators for Environmental Impact Assessments of Salmon Aquaculture." In: *Journal of Eukaryotic Microbiology*.

Fortuna, Miguel A., Daniel B. Stouffer, Jens M. Olesen, Pedro Jordano, David Mouillot, Boris R. Krasnov, Robert Poulin, and Jordi Bascompte (2010). "Nestedness versus modularity in ecological networks: two sides of the same coin?" In: *Journal of Animal Ecology* 79.4, pp. 811–817.

Fox, George E, Kenneth R Pechman, and Carl R Woese (1977). "Comparative cataloging of 16S ribosomal ribonucleic acid: molecular approach to procaryotic systematics." In: *International Journal of Systematic and Evolutionary Microbiology* 27.1, pp. 44–57.

Franzosa, Eric A, Lauren J McIver, Gholamali Rahnavard, Luke R Thompson, Melanie Schirmer, George Weingart, Karen Schwarzberg Lipson, Rob Knight, J Gregory Caporaso, Nicola Segata, and Curtis Huttenhower (2018). "Species-level functional profiling of metagenomes and metatranscriptomes." In: *Nat. Methods* 15.11, pp. 962–968.

Freilich, Mara A., Evie Wieters, Bernardo R. Broitman, Pablo A. Marquet, and Sergio A. Navarrete (2018). "Species co-occurrence networks: Can they reveal trophic and non-trophic interactions in ecological communities?" In: *Ecology* 99.3, pp. 690–699.

Friedman, Jonathan and Eric J Alm (2012). "Inferring correlation networks from genomic survey data." In: *PLoS Comput. Biol.* 8.9, e1002687.

Friedman, Nir, Michal Linial, Iftach Nachman, and Dana Pe' er (2000). "Using Bayesian Networks to Analyze Expression Data." In: *J. Comput. Biol.* 7.3-4, pp. 601–620.

Friedrich, Anja B, Isabell Fischer, Peter Proksch, Jörg Hacker, and Ute Hentschel (2001). "Temporal variation of the microbial community associated with the mediterranean sponge Aplysina aerophoba." In: *FEMS Microbiology Ecology* 38.2, pp. 105–113.

Fröstl, Jürgen M and Jörg Overmann (1998). "Physiology and tactic response of the phototrophic consortium "Chlorochromatium aggregatum"." In: *Archives of microbiology* 169.2, pp. 129–135.

Fu, Limin, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li (2012). "CD-HIT: accelerated for clustering the next-generation sequencing data." In: *Bioinformatics* 28.23, pp. 3150–3152.

García, Cristina, Manuel Rendueles, and Mario Díaz (2017). "Microbial amensalism in Lactobacillus casei and Pseudomonas taetrolens mixed culture." In: *Bioprocess and Biosystems Engineering* 40.7, pp. 1111–1122.

Garrity, George M. (2016). “A New Genomics-Driven Taxonomy of Bacteria and Archaea: Are We There Yet?” In: *Journal of Clinical Microbiology* 54.8, pp. 1956–1963.

Gause, G. F. (1936). “The Struggle for Existence.” In: *Soil Science* 41.2, p. 159.

Gavrish, Ekaterina, Annette Bollmann, Slava Epstein, and Kim Lewis (2008). “A trap for in situ cultivation of filamentous actinobacteria.” In: *Journal of Microbiological Methods* 72.3, pp. 257–262.

Geissinger, Oliver, Daniel P. R. Herlemann, Erhard Mörschel, Uwe G. Maier, and Andreas Brune (2009). “The Ultramicrobacterium “Elusimicrobium minutum”gen. nov., sp. nov., the First Cultivated Representative of the Termite Group 1 Phylum.” In: *Appl. Environ. Microbiol.* 75.9, pp. 2831–2840.

Gibbons, Sean M, Sean M Kearney, Chris S Smillie, and Eric J Alm (2017). “Two dynamic regimes in the human gut microbiome.” In: *PLoS Comput. Biol.* 13.2, e1005364.

Gich, Frederic, Monika Anna Janys, Melanie König, and Jörg Overmann (2012). “Enrichment of previously uncultured bacteria from natural complex communities by adhesion to solid surfaces.” In: *Environmental microbiology* 14.11, pp. 2984–2997.

Gifford, Scott M., Shalabh Sharma, Johanna M. Rinta-Kanto, and Mary Ann Moran (2011). “Quantitative analysis of a deeply sequenced marine microbial metatranscriptome.” In: *The ISME journal* 5.3, pp. 461–472.

Gilbert, Jack A, Joshua A Steele, J Gregory Caporaso, Lars Steinbrück, Jens Reeder, Ben Temperton, Susan Huse, Alice C McHardy, Rob Knight, Ian Joint, Paul Somerfield, Jed A Fuhrman, and Dawn Field (2012). “Defining seasonal marine microbial community dynamics.” In: *ISME J.* 6.2, pp. 298–308.

Gloor, Gregory B, Jean M Macklaim, Vera Pawlowsky-Glahn, and Juan J Egozcue (2017). “Microbiome datasets are compositional: and this is not optional.” In: *Frontiers in microbiology* 8, p. 2224.

Godfray, H. Charles J. (2002). “Challenges for taxonomy.” In: *Nature* 417.6884, pp. 17–19.

Gohl, Daryl M, Pajau Vangay, John Garbe, Allison MacLean, Adam Hauge, Aaron Becker, Trevor J Gould, Jonathan B Clayton, Timothy J Johnson, Ryan Hunter, Dan Knights, and Kenneth B Beckman (2016). “Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies.” In: *Nat. Biotechnol.* 34.9, pp. 942–949.

Goldford, Joshua E., Nanxi Lu, Djordje Bajić, Sylvie Estrela, Mikhail Tikhonov, Alicia Sanchez-Gorostiaga, Daniel Segrè, Pankaj Mehta, and Alvaro Sanchez (2018). “Emergent simplicity in microbial community assembly.” In: *Science* 361.6401, pp. 469–474.

Gonzalez, Antonio, Jose A Navas-Molina, Tomasz Kosciolek, Daniel McDonald, Yoshiki Vázquez-Baeza, Gail Ackermann, Jeff DeReus, Stefan Janssen, Austin D Swafford, Stephanie B Orchanian, Jon G Sanders, Joshua Shorenstein, Hannes Holste, Semar Petrus, Adam Robbins-Pianka, Colin J Brislawn, Mingxun Wang, Jai Ram Rideout, Evan Bolyen, Matthew Dillon, J Gregory Caporaso, Pieter C Dorrestein, and Rob Knight (2018). "Qiita: rapid, web-enabled microbiome meta-analysis." In: *Nat. Methods* 15.10, pp. 796–798.

Gonze, Didier, Leo Lahti, Jeroen Raes, and Karoline Faust (2017). "Multi-stability and the origin of microbial community types." In: *The ISME Journal* 11.10, pp. 2159–2166.

Goodrich, Julia K, Sara C Di Rienzi, Angela C Poole, Omry Koren, William A Walters, J Gregory Caporaso, Rob Knight, and Ruth E Ley (2014). "Conducting a Microbiome Study." In: *Cell* 158.2, pp. 250–262.

Goodwin, Sara, John D. McPherson, and W. Richard McCombie (2016). "Coming of age: ten years of next-generation sequencing technologies." In: *Nature Reviews Genetics* 17.6, pp. 333–351.

Goris, Johan, Konstantinos T Konstantinidis, Joel A Klappenbach, Tom Coenye, Peter Vandamme, and James M Tiedje (2007). "DNA-DNA hybridization values and their relationship to whole-genome sequence similarities." In: *Int. J. Syst. Evol. Microbiol.* 57.Pt 1, pp. 81–91.

Gotelli, Nicholas J. (2000). "Null Model Analysis of Species Co-Occurrence Patterns." In: *Ecology* 81.9, pp. 2606–2621.

Griffin, Ashleigh S., Stuart A. West, and Angus Buckling (2004). "Cooperation and competition in pathogenic bacteria." In: *Nature* 430.7003, pp. 1024–1027.

Grünberger, Alexander, Christopher Probst, Stefan Helfrich, Arun Nanda, Birgit Stute, Wolfgang Wiechert, Eric von Lieres, Katharina Nöh, Julia Frunzke, and Dietrich Kohlheyer (2015). "Spatiotemporal microbial single-cell analysis using a high-throughput microfluidics cultivation platform." In: *Cytometry Part A* 87.12, pp. 1101–1115.

Guerrero, Ricardo, Carlos Pedrós-Alió, Isabel Esteve, Jordi Mas, David Chase, and Lynn Margulis (1986). "Predatory prokaryotes: predation and primary consumption evolved in bacteria." In: *Proceedings of the National Academy of Sciences* 83.7, pp. 2138–2142.

Guimarães Jr, Paulo R., Mathias M. Pires, Pedro Jordano, Jordi Bascompte, and John N. Thompson (2017). "Indirect effects drive coevolution in mutualistic networks." In: *Nature* 550.7677, pp. 511–514.

Guy, Lionel and Thijs J G Ettema (2011). "The archaeal 'TACK'superphylum and the origin of eukaryotes." In: *Trends in microbiology* 19.12, pp. 580–7.

Hahn, Martin W. (2009). "Description of seven candidate species affiliated with the phylum Actinobacteria, representing planktonic freshwater bacteria." In: *International Journal of Systematic and Evolutionary Microbiology* 59.1, pp. 112–117.

HajiRassouliha, Amir, Andrew J Taberner, Martyn P Nash, and Poul M F Nielsen (2018). "Suitability of recent hardware accelerators (DSPs, FPGAs, and GPUs) for computer vision and image processing algorithms." In: *Signal Processing: Image Communication* 68, pp. 101–119.

Hajishengallis, George, Richard P. Darveau, and Michael A. Curtis (2012). "The keystone-pathogen hypothesis." In: *Nature Reviews. Microbiology* 10.10, pp. 717–725.

Hajishengallis, George, Shuang Liang, Mark A. Payne, Ahmed Hashim, Ravi Jotwani, Mehmet A. Eskan, Megan L. McIntosh, Asil Alsam, Keith L. Kirkwood, John D. Lambris, Richard P. Darveau, and Michael A. Curtis (2011). "Low-abundance biofilm species orchestrates inflammatory periodontal disease through the commensal microbiota and complement." In: *Cell Host & Microbe* 10.5, pp. 497–506.

Handelsman, Jo (2004). "Metagenomics: Application of Genomics to Uncultured Microorganisms." In: *Microbiology and Molecular Biology Reviews* 68.4, pp. 669–685.

Hansen, Susse Kirkelund, Paul B. Rainey, Janus A. J. Haagensen, and Søren Molin (2007). "Evolution of species interactions in a biofilm community." In: *Nature* 445.7127, pp. 533–536.

Haraldsson, Matilda, Mélanie Gerphagnon, Pauline Bazin, Jonathan Colombet, Samuele Tecchio, Télesphore Sime-Ngando, and Nathalie Niquil (2018). "Microbial parasites make cyanobacteria blooms less of a trophic dead end than commonly assumed." In: *The ISME Journal* 12.4, p. 1008.

Harcombe, William (2010). "Novel cooperation experimentally evolved between species." In: *Evolution; International Journal of Organic Evolution* 64.7, pp. 2166–2172.

Harrison, Peter W, Blaise Alako, Clara Amid, Ana Cerdeño-Tárraga, Iain Cleland, Sam Holt, Abdulrahman Hussein, Suran Jayathilaka, Simon Kay, Thomas Keane, et al. (2018). "The European Nucleotide Archive in 2018." In: *Nucleic acids research* 47.D1, pp. D84–D88.

Hasin, Yehudit, Marcus Seldin, and Aldons Lusis (2017). "Multi-omics approaches to disease." In: *Genome biology* 18.1, p. 83.

Hatzenpichler, Roland (2012). "Diversity, Physiology, and Niche Differentiation of Ammonia-Oxidizing Archaea." In: *Applied and Environmental Microbiology* 78.21, pp. 7501–7510.

Hawinkel, Stijn, Federico Mattiello, Luc Bijnens, and Olivier Thas (2017). "A broken promise: microbiome differential abundance methods do not control the false discovery rate." In: *Briefings in bioinformatics* 20.1, pp. 210–221.

Heckerman, David (1990). “Probabilistic similarity networks.” In: *Networks* 20.5, pp. 607–636.

Heckerman, David (1997). “Bayesian Networks for Data Mining.” In: *Data Mining and Knowledge Discovery* 1.1, pp. 79–119.

Heijden, Marcel G. A. van der, Roy Bakker, Joost Verwaal, Tanja R. Scheublin, Matthy Rutten, Richard van Logtestijn, and Christian Staehelin (2006). “Symbiotic bacteria as a determinant of plant community structure and plant productivity in dune grassland.” In: *FEMS microbiology ecology* 56.2, pp. 178–187.

Heijden, Marcel G. A. van der, John N. Klironomos, Margot Ursic, Peter Moutoglis, Ruth Streitwolf-Engel, Thomas Boller, Andres Wiemken, and Ian R. Sanders (1998). “Mycorrhizal fungal diversity determines plant biodiversity, ecosystem variability and productivity.” In: *Nature* 396.6706, pp. 69–72.

Heintz-Buschart, Anna and Paul Wilmes (2018). “Human Gut Microbiome: Function Matters.” In: *Trends Microbiol.* 26.7, pp. 563–574.

Hemarajata, Peera and James Versalovic (2013). “Effects of probiotics on gut microbiota: mechanisms of intestinal immunomodulation and neuromodulation.” In: *Therapeutic Advances in Gastroenterology* 6.1, pp. 39–51.

Hernández, Ester, Rafael Bargiela, María Suárez Diez, Anette Friedrichs, Ana Elena Pérez-Cobas, María José Gosalbes, Henrik Knecht, Mónica Martínez-Martínez, Jana Seifert, Martin von Bergen, Alejandro Artacho, Alicia Ruiz, Cristina Campoy, Amparo Latorre, Stephan J. Ott, Andrés Moya, Antonio Suárez, Vitor A. P. Martins dos Santos, and Manuel Ferrer (2013). “Functional consequences of microbial shifts in the human gastrointestinal tract linked to antibiotic treatment and obesity.” In: *Gut Microbes* 4.4, pp. 306–315.

Hey, Jody (2001). “The mind of the species problem.” In: *Trends Ecol. Evol.* 16.7, pp. 326–329.

Hibbing, Michael E., Clay Fuqua, Matthew R. Parsek, and S. Brook Peterson (2010). “Bacterial competition: surviving and thriving in the microbial jungle.” In: *Nature reviews. Microbiology* 8.1, pp. 15–25.

Hoegh-Guldberg, Ove (1999). “Climate change, coral bleaching and the future of the world's coral reefs.” In: *Marine and Freshwater Research* 50.8, pp. 839–866.

Hoek, Tim A., Kevin Axelrod, Tommaso Biancalani, Eugene A. Yurtsev, Jinghui Liu, and Jeff Gore (2016). “Resource Availability Modulates the Cooperative and Competitive Nature of a Microbial Cross-Feeding Mutualism.” In: *PLOS Biology* 14.8, e1002540.

Hoffman, Matthew D and Andrew Gelman (2014). “The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo.” In: *Journal of Machine Learning Research* 15.1, pp. 1593–1623.

Holmes, Elaine, Ruey Leng Loo, Jeremiah Stamler, Magda Bictash, Ivan KS Yap, Queenie Chan, Tim Ebbels, Maria De Iorio, Ian J Brown, Kirill A Veselkov, et al. (2008).

"Human metabolic phenotype diversity and its association with diet and blood pressure." In: *Nature* 453.7193, p. 396.

Holt, Kathryn E, Stephen C Watts, Michael Inouye, and Scott C Ritchie (Aug. 2018). "FastSpar: rapid and scalable correlation estimation for compositional data." In:

Hooper, Lora V., Tore Midtvedt, and Jeffrey I. Gordon (2002). "How host-microbial interactions shape the nutrient environment of the mammalian intestine." In: *Annual Review of Nutrition* 22, pp. 283–307.

Hooper, Lora V., Thaddeus S. Stappenbeck, Chieu V. Hong, and Jeffrey I. Gordon (2003). "Angiogenins: a new class of microbicidal proteins involved in innate immunity." In: *Nature Immunology* 4.3, pp. 269–273.

Houldcroft, Charlotte J., Mathew A. Beale, and Judith Breuer (2017). "Clinical and biological insights from viral genome sequencing." In: *Nature Reviews Microbiology* 15.3, pp. 183–192.

Huang, Chung-Yuan, Chuen-Tsai Sun, and Hsun-Cheng Lin (2005). "Influence of local information on social simulations in small-world network models." In: *Journal of Artificial Societies and Social Simulation* 8.4.

Hubbell, Stephen P. (2001). *The Unified Neutral Theory of Biodiversity and Biogeography (MPB-32)*. Princeton University Press.

Huber, Harald, Michael J. Hohn, Reinhard Rachel, Tanja Fuchs, Verena C. Wimmer, and Karl O. Stetter (2002). "A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont." In: *Nature* 417.6884, pp. 63–67.

Hugenholtz, Philip, Adam Skarshewski, and Donovan H Parks (2016). "Genome-based microbial taxonomy coming of age." In: *Cold Spring Harbor Perspectives in Biology* 8.6, a018085.

Hugerth, Luisa W and Anders F Andersson (2017). "Analysing microbial community composition through amplicon sequencing: from sampling to hypothesis testing." In: *Frontiers in microbiology* 8, p. 1561.

Huttenhower, Curtis, Dirk Gevers, Rob Knight, Sahar Abubucker, Jonathan H Badger, Asif T Chinwalla, Heather H Creasy, Ashlee M Earl, Michael G FitzGerald, Robert S Fulton, et al. (2012). "Structure, function and diversity of the healthy human microbiome." In: *Nature* 486.7402, p. 207.

Hutzil, Saúl, Santiago Sandoval-Motta, Alejandro Frank, and Maximino Aldana (2018). "Modeling the role of the microbiome in evolution." In: *Frontiers in physiology* 9, p. 1836.

Ingham, Colin J., Ad Sprenkels, Johan Bomer, Douwe Molenaar, Albert van den Berg, Johan E. T. van Hylckama Vlieg, and Willem M. de Vos (2007). "The micro-Petri dish, a million-well growth chip for the culture and high-throughput screening of microorganisms." In: *Proceedings of the National Academy of Sciences* 104.46, pp. 18217–18222.

Irlinger, Françoise and Jérôme Mounier (2009). “Microbial interactions in cheese: implications for cheese quality and safety.” In: *Current Opinion in Biotechnology* 20.2, pp. 142–148.

Isaksen, Mai Faurschou, Friedhelm Bak, and Bo Barker Jørgensen (1994). “Thermophilic sulfate-reducing bacteria in cold marine sediment.” In: *FEMS Microbiology Ecology* 14.1, pp. 1–8.

Jaccard, Paul (1901). “Étude comparative de la distribution florale dans une portion des Alpes et des Jura.” In: *Bull Soc Vaudoise Sci Nat* 37, pp. 547–579.

Jacquet, Claire, Charlotte Moritz, Lyne Morissette, Pierre Legagneux, François Massol, Philippe Archambault, and Dominique Gravel (2016). “No complexity-stability relationship in empirical ecosystems.” In: *Nature Communications* 7, p. 12573.

Johnson, Katerina V.-A. and Kevin R. Foster (2018). “Why does the microbiome affect behaviour?” In: *Nature Reviews Microbiology*, p. 1.

Jordán, Ferenc (2009). “Keystone species and food webs.” In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 364.1524, pp. 1733–1741.

Jordano, Pedro, Jordi Bascompte, and Jens M. Olesen (2003). “Invariant properties in coevolutionary networks of plant–animal interactions.” In: *Ecology Letters* 6.1, pp. 69–81.

Jousset, Alexandre, Christina Bienhold, Antonis Chatzinotas, Laure Gallien, Angélique Gobet, Viola Kurm, Kirsten Küsel, Matthias C Rillig, Damian W Rivett, Joana F Salles, Marcel G A van der Heijden, Noha H Youssef, Xiaowei Zhang, Zhong Wei, and W H Gera Hol (2017). “Where less may be more: how the rare biosphere pulls ecosystems strings.” In: *ISME J.* 11.4, pp. 853–862.

Kaeberlein, Tammi, Kim Lewis, and Slava S Epstein (2002). “Isolating" uncultivable" microorganisms in pure culture in a simulated natural environment.” In: *Science* 296.5570, pp. 1127–1129.

Kaiser, Dale and Richard Losick (1993). “How and why bacteria talk to each other.” In: *Cell* 73.5, pp. 873–885.

Kanehisa, Minoru, Miho Furumichi, Mao Tanabe, Yoko Sato, and Kanae Morishima (2017). “KEGG: new perspectives on genomes, pathways, diseases and drugs.” In: *Nucleic Acids Res.* 45.D1, pp. D353–D361.

Karlsson, Fredrik H, Valentina Tremaroli, Intawat Nookaew, Göran Bergström, Carl Johan Behre, Björn Fagerberg, Jens Nielsen, and Fredrik Bäckhed (2013). “Gut metagenome in European women with normal, impaired and diabetic glucose control.” In: *Nature* 498.7452, pp. 99–103.

Karst, Søren M., Morten S. Dueholm, Simon J. McIlroy, Rasmus H. Kirkegaard, Per H. Nielsen, and Mads Albertsen (2018). “Retrieval of a million high-quality, full-length microbial 16S and 18S rRNA gene sequences without primer bias.” In: *Nature Biotechnology* 36.2, pp. 190–195.

Kaul, Abhishek, Siddhartha Mandal, Ori Davidov, and Shyamal D Peddada (2017). "Analysis of Microbiome Data in the Presence of Excess Zeros." In: *Front. Microbiol.* 8, p. 2114.

Kaur, Harpreet and Ranjeet Singh (2011). "Two new species of Myxobolus (Myxozoa: Myxosporea: Bivalvulida) infecting an Indian major carp and a cat fish in wetlands of Punjab, India." In: *Journal of Parasitic Diseases: Official Organ of the Indian Society for Parasitology* 35.2, pp. 169–176.

Keegan, Kevin P, Elizabeth M Glass, and Folker Meyer (2016). "MG-RAST, a Metagenomics Service for Analysis of Microbial Community Structure and Function." In: *Methods Mol. Biol.* 1399, pp. 207–233.

Kennedy, Nicholas A., Alan W. Walker, Susan H. Berry, Sylvia H. Duncan, Freda M. Farquarson, Petra Louis, and John M. Thomson (2014). "The Impact of Different DNA Extraction Kits and Laboratories upon the Assessment of Human Gut Microbiota Composition by 16S rRNA Gene Sequencing." In: *PLOS ONE* 9.2, e88982.

Kenters, Nikki, Gemma Henderson, Jeyamalar Jeyanathan, Sandra Kittelmann, and Peter H. Janssen (2011). "Isolation of previously uncultured rumen bacteria by dilution to extinction using a new liquid culture medium." In: *Journal of Microbiological Methods* 84.1, pp. 52–60.

Kimura, Motoo (1983). *The neutral theory of molecular evolution*. Cambridge University Press.

Kleiner, Manuel, Cecilia Wentrup, Christian Lott, Hanno Teeling, Silke Wetzel, Jacque Young, Yun-Juan Chang, Manesh Shah, Nathan C. VerBerkmoes, Jan Zarzycki, Georg Fuchs, Stephanie Markert, Kristina Hempel, Birgit Voigt, Dörte Becher, Manuel Liebeke, Michael Lalk, Dirk Albrecht, Michael Hecker, Thomas Schweder, and Nicole Dubilier (2012). "Metaproteomics of a gutless marine worm and its symbiotic microbial community reveal unusual pathways for carbon and energy use." In: *Proceedings of the National Academy of Sciences* 109.19, E1173–E1182.

Klindworth, Anna, Elmar Pruesse, Timmy Schweer, Jörg Peplies, Christian Quast, Matthias Horn, and Frank Oliver Glöckner (2013). "Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies." In: *Nucleic acids research* 41.1, e1–e1.

Klug, A. (1968). "Rosalind Franklin and the Discovery of the Structure of DNA." In: *Nature* 219.5156, pp. 808–844.

Knight, Rob, Alison Vrbanac, Bryn C Taylor, Alexander Aksenov, Chris Callewaert, Justine Debelius, Antonio Gonzalez, Tomasz Kosciolek, Laura-Isobel McCall, Daniel McDonald, Alexey V Melnik, James T Morton, Jose Navas, Robert A Quinn, Jon G Sanders, Austin D Swafford, Luke R Thompson, Anupriya Tripathi, Zhenjiang Z Xu, Jesse R Zaneveld, Qiyun Zhu, J Gregory Caporaso, and Pieter C Dorrestein

(2018). “Best practices for analysing microbiomes.” In: *Nat. Rev. Microbiol.* 16.7, pp. 410–422.

Knights, Dan, Elizabeth K Costello, and Rob Knight (2011a). “Supervised classification of human microbiota.” In: *FEMS Microbiol. Rev.* 35.2, pp. 343–359.

Knights, Dan, Justin Kuczynski, Emily S Charlson, Jesse Zaneveld, Michael C Mozer, Ronald G Collman, Frederic D Bushman, Rob Knight, and Scott T Kelley (2011b). “Bayesian community-wide culture-independent microbial source tracking.” In: *Nat. Methods* 8.9, pp. 761–763.

Knights, Dan, Justin Kuczynski, Omry Koren, Ruth E Ley, Dawn Field, Rob Knight, Todd Z DeSantis, and Scott T Kelley (2011c). “Supervised classification of microbiota mitigates mislabeling errors.” In: *ISME J.* 5.4, pp. 570–573.

Knights, Dan, Tonya L Ward, Christopher E McKinlay, Hannah Miller, Antonio Gonzalez, Daniel McDonald, and Rob Knight (2014). “Rethinking “enterotypes”.” In: *Cell host & microbe* 16.4, pp. 433–437.

Kodratoff, Yves and Ryszard S Michalski (2014). *Machine Learning: An Artificial Intelligence Approach*. Elsevier.

Kokkoris, Giorgos D., Vincent A. A. Jansen, Michel Loreau, and Andreas Y. Troumbis (2002). “Variability in interaction strength and implications for biodiversity.” In: *Journal of Animal Ecology* 71.2, pp. 362–371.

Koller, Daphne, Nir Friedman, and Francis Bach (2009). *Probabilistic Graphical Models: Principles and Techniques*. MIT Press.

Kolter, Roberto and E. Peter Greenberg (2006). “Microbial sciences: The superficial life of microbes.” In: *Nature* 441, pp. 300–302.

Könneke, Martin, Anne E. Bernhard, José R. de la Torre, Christopher B. Walker, John B. Waterbury, and David A. Stahl (2005). “Isolation of an autotrophic ammonia-oxidizing marine archaeon.” In: *Nature* 437.7058, pp. 543–546.

Kumar, S., K. G. Mukerji, and R. Lal (1996). “Molecular aspects of pesticide degradation by microorganisms.” In: *Critical Reviews in Microbiology* 22.1, pp. 1–26.

Kunin, Victor, Anna Engelbrektson, Howard Ochman, and Philip Hugenholtz (2010). “Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates.” In: *Environ. Microbiol.* 12.1, pp. 118–123.

Kurtz, Zachary D, Christian L Müller, Emily R Miraldi, Dan R Littman, Martin J Blaser, and Richard A Bonneau (2015). “Sparse and compositionally robust inference of microbial ecological networks.” In: *PLoS computational biology* 11.5, e1004226.

Laforest-Lapointe, Isabelle and Marie-Claire Arrieta (2018). “Microbial Eukaryotes: a Missing Link in Gut Microbiome Studies.” In: *mSystems* 3.2, e00201–17.

Lage, Olga Maria and Joana Bondoso (2012). “Bringing Planctomycetes into pure culture.” In: *Frontiers in microbiology* 3, p. 405.

Lagier, Jean-Christophe, Grégory Dubourg, Matthieu Million, Frédéric Cadoret, Melhem Bilen, Florence Fenollar, Anthony Levasseur, Jean-Marc Rolain, Pierre-Edouard Fournier, and Didier Raoult (2018). "Culturing the human microbiota and culturomics." In: *Nature Reviews Microbiology* 16.9, p. 540.

Lagkouvardos, Ilias, Divya Joseph, Martin Kapfhammer, Sabahattin Giritli, Matthias Horn, Dirk Haller, and Thomas Clavel (2016). "IMNGS: A comprehensive open resource of processed 16S rRNA microbial profiles for ecology and diversity studies." In: *Scientific reports* 6, p. 33721.

Lane, Nick and William Martin (2010). "The energetics of genome complexity." In: *Nature* 467.7318, pp. 929–934.

Lapage, Stephen P, Peter HA Sneath, Erwin F Lessel, VBD Skerman, HPR Seeliger, and WA Clark (1992). *International code of nomenclature of bacteria: bacteriological code, 1990 revision*. ASM Press.

Law, Charity W, Yunshun Chen, Wei Shi, and Gordon K Smyth (2014). "voom: Precision weights unlock linear model analysis tools for RNA-seq read counts." In: *Genome Biology* 15.2, R29.

Lawley, Trevor D., Donna M. Bouley, Yana E. Hoy, Christine Gerke, David A. Relman, and Denise M. Monack (2008). "Host Transmission of Salmonella enterica Serovar Typhimurium Is Controlled by Virulence Factors and Indigenous Intestinal Microbiota." In: *Infection and Immunity* 76.1, pp. 403–416.

Lawrence, Jeffrey G. and Adam C. Retchless (2009). "The interplay of homologous recombination and horizontal gene transfer in bacterial speciation." In: *Methods in Molecular Biology (Clifton, N.J.)* 532, pp. 29–53.

Layeghifard, Mehdi, David M. Hwang, and David S. Guttman (2017). "Disentangling Interactions in the Microbiome: A Network Perspective." In: *Trends in Microbiology* 25.3, pp. 217–228.

Leek, Jeffrey T, Robert B Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A Irizarry (2010). "Tackling the widespread and critical impact of batch effects in high-throughput data." In: *Nat. Rev. Genet.* 11.10, pp. 733–739.

Leek, Jeffrey T and John D Storey (2007). "Capturing heterogeneity in gene expression studies by surrogate variable analysis." In: *PLoS Genet.* 3.9, pp. 1724–1735.

Leigh, E. G. (2007). "Neutral theory: a historical perspective." In: *Journal of Evolutionary Biology* 20.6, pp. 2075–2091.

Leinonen, Rasko, Hideaki Sugawara, Martin Shumway, and International Nucleotide Sequence Database Collaboration (2010). "The sequence read archive." In: *Nucleic acids research* 39, pp. D19–D21.

Lema, Kimberley A., Bette L. Willis, and David G. Bourne (2012). "Corals Form Characteristic Associations with Symbiotic Nitrogen-Fixing Bacteria." In: *Appl. Environ. Microbiol.* 78.9, pp. 3136–3144.

Lemos, M. L., C. P. Dopazo, A. E. Toranzo, and J. L. Barja (1991). "Competitive dominance of antibiotic-producing marine bacteria in mixed cultures." In: *The Journal of Applied Bacteriology* 71.3, pp. 228–232.

Leschine, S. B. (1995). "Cellulose degradation in anaerobic environments." In: *Annual Review of Microbiology* 49, pp. 399–426.

Lesser, Michael P., Charles H. Mazel, Maxim Y. Gorbunov, and Paul G. Falkowski (2004). "Discovery of Symbiotic Nitrogen-Fixing Cyanobacteria in Corals." In: *Science* 305.5686, pp. 997–1000.

Levy, Asaf, Jonathan M. Conway, Jeffery L. Dangl, and Tanja Woyke (2018). "Elucidating Bacterial Gene Functions in the Plant Microbiome." In: *Cell Host & Microbe* 24.4, pp. 475–485.

Levy, Roie and Elhanan Borenstein (2012). "Reverse Ecology: from systems to environments and back." In: *Adv. Exp. Med. Biol.* 751, pp. 329–345.

Levy, Roie and Elhanan Borenstein (2013). "Metabolic modeling of species interaction in the human microbiome elucidates community-level assembly rules." In: *Proceedings of the National Academy of Sciences* 110.31, pp. 12804–12809.

Ley, Ruth E., Micah Hamady, Catherine Lozupone, Peter Turnbaugh, Rob Roy Ramey, J. Stephen Bircher, Michael. L. Schlegel, Tammy A. Tucker, Mark D. Schrenzel, Rob Knight, and Jeffrey I. Gordon (2008). "Evolution of mammals and their gut microbes." In: *Science (New York, N.Y.)* 320.5883, pp. 1647–1651.

Li, Simone S, Ana Zhu, Vladimir Benes, Paul I Costea, Rajna Hercog, Falk Hildebrand, Jaime Huerta-Cepas, Max Nieuwdorp, Jarkko Salojärvi, Anita Y Voigt, Georg Zeller, Shinichi Sunagawa, Willem M de Vos, and Peer Bork (2016). "Durable coexistence of donor and recipient strains after fecal microbiota transplantation." In: *Science* 352.6285, pp. 586–589.

Liao, Jingqiu, Xiaofeng Cao, Lei Zhao, Jie Wang, Zhe Gao, Michael Cai Wang, and Yi Huang (2016). "The importance of neutral and niche processes for bacterial community assembly differs between habitat generalists and specialists." In: *FEMS microbiology ecology* 92.11.

Lidicker, William Z. (1979). "A Clarification of Interactions in Ecological Systems." In: *BioScience* 29.8, pp. 475–477.

Lima-Mendez, Gipsi, Karoline Faust, Nicolas Henry, Johan Decelle, Sébastien Colin, Fabrizio Carcillo, Samuel Chaffron, J Cesar Ignacio-Espinosa, Simon Roux, Flora Vincent, Lucie Bittner, Youssef Darzi, Jun Wang, Stéphane Audic, Léo Berline, Gianluca Bontempi, Ana M Cabello, Laurent Coppola, Francisco M Cornejo-Castillo, Francesco d'Ovidio, Luc De Meester, Isabel Ferrera, Marie-José Garet-Delmas, Lionel

Guidi, Elena Lara, Stéphane Pesant, Marta Royo-Llonch, Guillem Salazar, Pablo Sánchez, Marta Sebastian, Caroline Souffreau, Céline Dimier, Marc Picheral, Sarah Searson, Stefanie Kandels-Lewis, Tara Oceans coordinators, Gabriel Gorsky, Fabrice Not, Hiroyuki Ogata, Sabrina Speich, Lars Stemmann, Jean Weissenbach, Patrick Wincker, Silvia G Acinas, Shinichi Sunagawa, Peer Bork, Matthew B Sullivan, Eric Karsenti, Chris Bowler, Colomban de Vargas, and Jeroen Raes (2015). "Ocean plankton. Determinants of community structure in the global plankton interactome." In: *Science* 348.6237, p. 1262073.

Lima-Mendez, Gipsi and Jacques van Helden (2009). "The powerful law of the power law and other myths in network biology." In: *Molecular bioSystems* 5.12, pp. 1482–1493.

Liu, Angela, Anne M Archer, Matthew B Biggs, and Jason A Papin (2017). "Growth-altering microbial interactions are responsive to chemical context." In: *PloS one* 12.3, e0164919.

Lo, Chieh and Radu Marculescu (2017). "MPLasso: Inferring microbial association networks using prior microbial knowledge." In: *PLoS Comput. Biol.* 13.12, e1005915.

Locey, Kenneth J. and Jay T. Lennon (2016). "Scaling laws predict global microbial diversity." In: *Proceedings of the National Academy of Sciences*, p. 201521291.

Login, Frédéric H., Séverine Balmand, Agnès Vallier, Carole Vincent-Monégat, Aurélien Vigneron, Michèle Weiss-Gayet, Didier Rochat, and Abdelaziz Heddi (2011). "Antimicrobial Peptides Keep Insect Endosymbionts Under Control." In: *Science* 334.6054, pp. 362–365.

Lombardo, Michael P. (2008). "Access to mutualistic endosymbiotic microbes: an underappreciated benefit of group living." In: *Behavioral Ecology and Sociobiology* 62.4, pp. 479–497.

Long, Richard A., Damien Eveillard, Shelli L. M. Franco, Eric Reeves, and James L. Pinckney (2013). "Antagonistic interactions between heterotrophic bacteria as a potential regulator of community structure of hypersaline microbial mats." In: *FEMS Microbiology Ecology* 83.1, pp. 74–81.

Lopez-Siles, Mireia, Sylvia H Duncan, L Jesús Garcia-Gil, and Margarita Martinez-Medina (2017). "Faecalibacterium prausnitzii: from microbiology to diagnostics and prognostics." In: *The ISME journal* 11.4, pp. 841–852.

Louca, Stilianos, Martin F. Polz, Florent Mazel, Michaeline B. N. Albright, Julie A. Huber, Mary I. O'Connor, Martin Ackermann, Aria S. Hahn, Diane S. Srivastava, Sean A. Crowe, Michael Doebeli, and Laura Wegener Parfrey (2018). "Function and functional redundancy in microbial systems." In: *Nature Ecology & Evolution* 2.6, pp. 936–943.

Love, Michael I, Wolfgang Huber, and Simon Anders (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." In: *Genome Biol.* 15.12, p. 550.

Lozupone, Catherine and Rob Knight (2005). "UniFrac: a new phylogenetic method for comparing microbial communities." In: *Appl. Environ. Microbiol.* 71.12, pp. 8228–8235.

Lozupone, Catherine A, Micah Hamady, Scott T Kelley, and Rob Knight (2007). "Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities." In: *Appl. Environ. Microbiol.* 73.5, pp. 1576–1585.

Lozupone, Catherine A and Rob Knight (2008). "Species divergence and the measurement of microbial diversity." In: *FEMS Microbiol. Rev.* 32.4, pp. 557–578.

Lu, Min, Jiri Hulcr, and Jianghua Sun (2016). "The Role of Symbiotic Microbes in Insect Invasions." In: *Annual Review of Ecology, Evolution, and Systematics* 47.1, pp. 487–505.

Luscombe, Nicholas M, M Madan Babu, Haiyuan Yu, Michael Snyder, Sarah A Teichmann, and Mark Gerstein (2004). "Genomic analysis of regulatory network dynamics reveals large topological changes." In: *Nature* 431.7006, p. 308.

Lynch, Michael D J and Josh D Neufeld (2015). "Ecology and exploration of the rare biosphere." In: *Nat. Rev. Microbiol.* 13.4, pp. 217–229.

Lyons, Nicholas A. and Roberto Kolter (2015). "On The Evolution of Bacterial Multicellularity." In: *Current opinion in microbiology* 24, pp. 21–28.

Lyte, Mark (2013). "Microbial endocrinology in the microbiome-gut-brain axis: how bacterial production and utilization of neurochemicals influence behavior." In: *PLoS pathogens* 9.11, e1003726.

Ma, Yonghui, Hua Chen, Canhui Lan, and Jianlin Ren (2018). "Help, hope and hype: ethical considerations of human microbiome research and applications." In: *Protein & Cell* 9.5, pp. 404–415.

Macfarlane, Sandra and George T. Macfarlane (2003). "Regulation of short-chain fatty acid production." In: *Proceedings of the Nutrition Society* 62.1, pp. 67–72.

Madigan, Michael T., John M. Martinko, Kelly S. Bender, Daniel H. Buckley, David A. Stahl, and Thomas Brock (2014). *Brock Biology of Microorganisms, 14th edition*. Pearson.

Mahé, Frédéric, Torbjørn Rognes, Christopher Quince, Colomban de Vargas, and Micah Dunthorn (2014). "Swarm: robust and fast clustering method for amplicon-based studies." In: *PeerJ* 2, e593.

Mahowald, Michael A., Federico E. Rey, Henning Seedorf, Peter J. Turnbaugh, Robert S. Fulton, Aye Wollam, Neha Shah, Chunyan Wang, Vincent Magrini, Richard K. Wilson, Brandi L. Cantarel, Pedro M. Coutinho, Bernard Henrissat, Lara W. Crock,

Alison Russell, Nathan C. Verberkmoes, Robert L. Hettich, and Jeffrey I. Gordon (2009). "Characterizing a model human gut microbiota composed of members of its two dominant bacterial phyla." In: *Proceedings of the National Academy of Sciences of the United States of America* 106.14, pp. 5859–5864.

Mandal, Siddhartha, Will Van Treuren, Richard A White, Merete Eggesbø, Rob Knight, and Shyamal D Peddada (2015). "Analysis of composition of microbiomes: a novel method for studying microbial composition." In: *Microb. Ecol. Health Dis.* 26, p. 27663.

Mandel, Mark J., Michael S. Wollenberg, Eric V. Stabb, Karen L. Visick, and Edward G. Ruby (2009). "A single regulatory gene is sufficient to alter bacterial host range." In: *Nature* 458.7235, pp. 215–218.

Mark Welch, Jessica L., Blair J. Rossetti, Christopher W. Rieken, Floyd E. Dewhirst, and Gary G. Borisy (2016). "Biogeography of a human oral microbiome at the micron scale." In: *Proceedings of the National Academy of Sciences of the United States of America* 113.6, E791–800.

Martens, Eric C., Nicole M. Koropatkin, Thomas J. Smith, and Jeffrey I. Gordon (2009). "Complex Glycan Catabolism by the Human Gut Microbiota: The Bacteroidetes Sus-like Paradigm." In: *The Journal of Biological Chemistry* 284.37, pp. 24673–24677.

Martijn, Joran and Thijs J G Ettema (2013). "From archaeon to eukaryote: the evolutionary dark ages of the eukaryotic cell." In: *Biochemical Society transactions* 41.1, pp. 451–7.

Martin, Andrew P (2002). "Phylogenetic approaches for describing and comparing the diversity of microbial communities." In: *Appl. Environ. Microbiol.* 68.8, pp. 3673–3682.

Martin, François-Pierre J, Marc-Emmanuel Dumas, Yulan Wang, Cristina Legido-Quigley, Ivan KS Yap, Huiru Tang, Séverine Zirah, Gerard M Murphy, Olivier Cloarec, John C Lindon, et al. (2007). "A top-down systems biology view of microbiome-mammalian metabolic interactions in a mouse model." In: *Molecular systems biology* 3.1, p. 112.

Martín-Fernández, JA and S Thió-Henestrosa (2006). "Rounded zeros: some practical aspects for compositional data." In: *Geological Society, London, Special Publications* 264.1, pp. 191–201.

Martín-Fernández, Josep A, Carles Barceló-Vidal, and Vera Pawlowsky-Glahn (2003). "Dealing with zeros and missing values in compositional data sets using nonparametric imputation." In: *Mathematical Geology* 35.3, pp. 253–278.

Mashburn, Lauren M. and Marvin Whiteley (2005). "Membrane vesicles traffic signals and facilitate group activities in a prokaryote." In: *Nature* 437.7057, pp. 422–425.

Matias Rodrigues, João F and Christian von Mering (2014). “HPC-CLUST: distributed hierarchical clustering for large sets of nucleotide sequences.” In: *Bioinformatics* 30.2, pp. 287–288.

Matias Rodrigues, João F, Thomas S B Schmidt, Janko Tackmann, and Christian von Mering (2017). “MAPseq: highly efficient k-mer search with confidence estimates, for rRNA sequence analysis.” In: *Bioinformatics* 33.23, pp. 3808–3810.

May, Robert M. (1972). “Will a Large Complex System be Stable?” In: *Nature* 238.5364, pp. 413–414.

May, Robert McCredie (2001). *Stability and complexity in model ecosystems*. Vol. 6. Princeton university press.

Mayer, Emeran A. (2011). “Gut feelings: the emerging biology of gut-brain communication.” In: *Nature Reviews. Neuroscience* 12.8, pp. 453–466.

Mayr, Ernest (1942). “Systematics and the origin of species.” In: *Columbia Univ. Press, New York*.

Mazmanian, Sarkis K, Hua Liu Cui, Arthur O. Tzianabos, and Dennis L. Kasper (2005). “An immunomodulatory molecule of symbiotic bacteria directs maturation of the host immune system.” In: *Cell* 122.1, pp. 107–118.

McCann, K. S. (2000). “The diversity-stability debate.” In: *Nature* 405.6783, pp. 228–233.

McDonald, Daniel, Embriette Hyde, Justine W Debelius, James T Morton, Antonio Gonzalez, Gail Ackermann, Alexander A Aksenov, Bahar Behsaz, Caitriona Brennan, Yingfeng Chen, et al. (2018). “American gut: an open platform for citizen science microbiome research.” In: *mSystems* 3.3, e00031–18.

McFall-Ngai, Margaret (2007). “Adaptive Immunity: Care for the community.” In: *Nature* 445, p. 153.

McFall-Ngai, Margaret (2008). “Are biologists in 'future shock'? Symbiosis integrates biology across domains.” In: *Nature Reviews Microbiology* 6.10, pp. 789–792.

McFall-Ngai, Margaret, Michael G. Hadfield, Thomas C. G. Bosch, Hannah V. Carey, Tomislav Domazet-Lošo, Angela E. Douglas, Nicole Dubilier, Gerard Eberl, Tadashi Fukami, Scott F. Gilbert, Ute Hentschel, Nicole King, Staffan Kjelleberg, Andrew H. Knoll, Natacha Kremer, Sarkis K. Mazmanian, Jessica L. Metcalf, Kenneth Nealson, Naomi E. Pierce, John F. Rawls, Ann Reid, Edward G. Ruby, Mary Rumpho, Jon G. Sanders, Diethard Tautz, and Jennifer J. Wernegreen (2013). “Animals in a bacterial world, a new imperative for the life sciences.” In: *Proceedings of the National Academy of Sciences* 110.9, pp. 3229–3236.

McFall-Ngai, Margaret J. and Edward G. Ruby (1998). “Sepiolids and Vibrios: When First They Meet.” In: *BioScience* 48.4, pp. 257–265.

McFarland, Lynne V. (2008). “Antibiotic-associated diarrhea: epidemiology, trends and treatment.” In: *Future Microbiology* 3.5, pp. 563–578.

McGeachie, Michael J, Joanne E Sordillo, Travis Gibson, George M Weinstock, Yang-Yu Liu, Diane R Gold, Scott T Weiss, and Augusto Litonjua (2016). "Longitudinal Prediction of the Infant Gut Microbiome with Dynamic Bayesian Networks." In: *Sci. Rep.* 6, p. 20359.

McGill, Brian J., Brian A. Maurer, and Michael D. Weiser (2006). "Empirical Evaluation of Neutral Theory." In: *Ecology* 87.6, pp. 1411–1423.

McIver, Lauren J, Galeb Abu-Ali, Eric A Franzosa, Randall Schwager, Xochitl C Morgan, Levi Waldron, Nicola Segata, and Curtis Huttenhower (2017). "bioBakery: a meta'omic analysis environment." In: *Bioinformatics* 34.7, pp. 1235–1237.

McKenney, Peter T. and Eric G. Pamer (2015). "From hype to hope: the gut microbiota in enteric infectious disease." In: *Cell* 163.6, pp. 1326–1332.

McMurdie, Paul J and Susan Holmes (2014). "Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible." In: *PLoS Comput. Biol.* 10.4, e1003531.

Medini, Duccio, Davide Serruto, Julian Parkhill, David A. Relman, Claudio Donati, Richard Moxon, Stanley Falkow, and Rino Rappuoli (2008). "Microbiology in the post-genomic era." In: *Nature Reviews Microbiology* 6.6, pp. 419–430.

Meinesz, A, L Benichou, J Blachier, T Komatsu, R Lemée, H Molenaar, and X Mari (1995). "Variations in the structure, morphology and biomass of Caulerpa taxifolia in the Mediterranean Sea." In: *Botanica marina* 38.1-6, pp. 499–508.

Mendes-Soares, Helena, Michael Mundy, Luis Mendes Soares, and Nicholas Chia (2016). "MMinte: an application for predicting metabolic interactions among the microbial species in a community." In: *BMC Bioinformatics* 17.1, p. 343.

Menge, Bruce A. (1995). "Indirect Effects in Marine Rocky Intertidal Interaction Webs: Patterns and Importance." In: *Ecological Monographs* 65.1, pp. 21–74.

Menon, Rajita, Vivek Ramanan, and Kirill S Korolev (2018). "Interactions between species introduce spurious associations in microbiome studies." In: *PLoS Comput. Biol.* 14.1, e1005939.

Meyer-Hoffert, U., M. W. Hornef, B. Henriques-Normark, L.-G. Axelsson, T. Midtvedt, K. Pütsep, and M. Andersson (2008). "Secreted enteric antimicrobial activity localises to the mucus surface layer." In: *Gut* 57.6, pp. 764–771.

Mills, L. Scott, Michael E. Soulé, and Daniel F. Doak (1993). "The Keystone-Species Concept in Ecology and ConservationManagement and policy must explicitly consider the complexity of interactions in natural systems." In: *BioScience* 43.4, pp. 219–224.

Milns, Isobel, Colin M Beale, and V Anne Smith (2010). "Revealing ecological networks using Bayesian network inference algorithms." In: *Ecology* 91.7, pp. 1892–1899.

Mitchell, Alex L, Maxim Scheremetjew, Hubert Denise, Simon Potter, Aleksandra Tarkowska, Matloob Qureshi, Gustavo A Salazar, Sebastien Pesseat, Miguel A Boland, Fiona M I Hunter, Petra Ten Hoopen, Blaise Alako, Clara Amid, Darren J Wilkinson,

Thomas P Curtis, Guy Cochrane, and Robert D Finn (2018). “EBI Metagenomics in 2017: enriching the analysis of microbial communities, from sequence reads to assemblies.” In: *Nucleic Acids Res.* 46.D1, pp. D726–D735.

Mitra, Suparna, Paul Rupek, Daniel C Richter, Tim Urich, Jack A Gilbert, Folker Meyer, Andreas Wilke, and Daniel H Huson (2011). “Functional analysis of metagenomes and metatranscriptomes using SEED and KEGG.” In: *BMC Bioinformatics* 12.Suppl 1, S21.

Mitton, Jeffry B. and Michael C. Grant (1996). “Genetic Variation and the Natural History of Quaking AspenThe ways in which aspen reproduces underlie its great geographic range, high levels of genetic variability, and persistence.” In: *BioScience* 46.1, pp. 25–31.

Montoya, José M., Stuart L. Pimm, and Ricard V. Solé (2006). “Ecological networks and their fragility.” In: *Nature* 442.7100, pp. 259–264.

Montoya, JOSE M. and RICARD V. Solé (2002). “Small World Patterns in Food Webs.” In: *Journal of Theoretical Biology* 214.3, pp. 405–412.

Moore, Peter B (1999). “A biophysical chemist’s thoughts on cell size.” In: *Size limits of very small microorganisms. National Academic Press, Washington DC*, pp. 16–20.

Mormile, Melanie R., Bo-Young Hong, and Kathleen C. Benison (2009). “Molecular Analysis of the Microbial Communities of Mars Analog Lakes in Western Australia.” In: *Astrobiology* 9.10, pp. 919–930.

Morris, Robert M., Michael S. Rappé, Stephanie A. Connon, Kevin L. Vergin, William A. Siebold, Craig A. Carlson, and Stephen J. Giovannoni (2002). “SAR11 clade dominates ocean surface bacterioplankton communities.” In: *Nature* 420.6917, pp. 806–810.

Mosca, Alexis, Marion Leclerc, and Jean P Hugot (2016). “Gut Microbiota Diversity and Human Diseases: Should We Reintroduce Key Predators in Our Ecosystem?” In: *Front. Microbiol.* 7, p. 455.

Moult, John, Krzysztof Fidelis, Andriy Kryshtafovych, Torsten Schwede, and Anna Tramontano (2016). “Critical assessment of methods of protein structure prediction: Progress and new directions in round XI.” In: *Proteins: Structure, Function, and Bioinformatics* 84, pp. 4–14.

Mounier, Jérôme, Christophe Monnet, Tatiana Vallaeys, Roger Arditi, Anne-Sophie Sarthou, Arnaud Hélias, and Françoise Irlinger (2008). “Microbial Interactions within a Cheese Microbial Community.” In: *Applied and Environmental Microbiology* 74.1, pp. 172–181.

Moya, Andrés and Manuel Ferrer (2016). “Functional Redundancy-Induced Stability of Gut Microbiota Subjected to Disturbance.” In: *Trends in Microbiology* 24.5, pp. 402–413.

Muegge, Brian D, Justin Kuczynski, Dan Knights, Jose C Clemente, Antonio González, Luigi Fontana, Bernard Henrissat, Rob Knight, and Jeffrey I Gordon (2011). "Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans." In: *Science* 332.6032, pp. 970–974.

Mukherjee, S, S Vaishnava, and LV Hooper (2008). "Multi-layered regulation of intestinal antimicrobial defense." In: *Cellular and Molecular Life Sciences* 65.19, pp. 3019–3027.

Murray, RGE and E Stackebrandt (1995). "Taxonomic note: implementation of the provisional status Candidatus for incompletely described procaryotes." In: *International Journal of Systematic and Evolutionary Microbiology* 45.1, pp. 186–187.

Nakatsuji, Teruaki, Hsin-I. Chiang, Shangi B. Jiang, Harish Nagarajan, Karsten Zengler, and Richard L. Gallo (2013). "The microbiome extends to subepidermal compartments of normal skin." In: *Nature Communications* 4, p. 1431.

Narendra, Varun, Nikita I Lytkin, Constantin F Aliferis, and Alexander Statnikov (2011). "A comprehensive assessment of methods for de-novo reverse-engineering of genome-scale regulatory networks." In: *Genomics* 97.1, pp. 7–18.

Neal, Radford M et al. (2011). "MCMC using Hamiltonian dynamics." In: *Handbook of markov chain monte carlo* 2.11, p. 2.

Needham, David M, Rohan Sachdeva, and Jed A Fuhrman (2017). "Ecological dynamics and co-occurrence among marine phytoplankton, bacteria and myoviruses shows microdiversity matters." In: *ISME J.* 11.7, pp. 1614–1629.

Newman, M. (2003). "The Structure and Function of Complex Networks." In: *SIAM Review* 45.2, pp. 167–256.

Nichols, D, N Cahoon, EM Trakhtenberg, L Pham, A Mehta, A Belanger, T Kanigan, K Lewis, and SS Epstein (2010). "Use of ichip for high-throughput in situ cultivation of "uncultivable" microbial species." In: *Appl. Environ. Microbiol.* 76.8, pp. 2445–2450.

Nilsson, L. Anders (1988). "The evolution of flowers with deep corolla tubes." In: *Nature* 334.6178, pp. 147–149.

Nishida, Atsushi, Ryo Inoue, Osamu Inatomi, Shigeki Bamba, Yuji Naito, and Akira Andoh (2018). "Gut microbiota in the pathogenesis of inflammatory bowel disease." In: *Clinical Journal of Gastroenterology* 11.1, pp. 1–10.

Nowak, Bożena, Jolanta Pająk, Magdalena Drozd-Bratkowicz, and Grażyna Rymarz (2011). "Microorganisms participating in the biodegradation of modified polyethylene films in different soils under laboratory conditions." In: *International Biodeterioration & Biodegradation* 65.6, pp. 757–767.

Oberhardt, Matthew A., Raphy Zarecki, Sabine Gronow, Elke Lang, Hans-Peter Klenk, Uri Gophna, and Eytan Ruppin (2015). "Harnessing the landscape of microbial culture media to predict new organism–media pairings." In: *Nature Communications* 6, p. 8493.

Oh, Phaik Lyn, Andrew K Benson, Daniel A Peterson, Prabhu B Patil, Etsuko N Moriyama, Stefan Roos, and Jens Walter (2010). "Diversification of the gut symbiont Lactobacillus reuteri as a result of host-driven evolution." In: *The ISME journal* 4.3, p. 377.

Oliver, Tom H., Matthew S. Heard, Nick J. B. Isaac, David B. Roy, Deborah Procter, Felix Eigenbrod, Rob Freckleton, Andy Hector, C. David L. Orme, Owen L. Petchey, Vânia Proença, David Raffaelli, K. Blake Suttle, Georgina M. Mace, Berta Martín-López, Ben A. Woodcock, and James M. Bullock (2015). "Biodiversity and Resilience of Ecosystem Functions." In: *Trends in Ecology & Evolution* 30.11, pp. 673–684.

Olson, Nathan D, Todd J Treangen, Christopher M Hill, Victoria Cepeda-Espinoza, Jay Ghurye, Sergey Koren, and Mihai Pop (2017). "Metagenomic assembly through the lens of validation: recent advances in assessing and improving the quality of genomes assembled from metagenomes." In: *Briefings in bioinformatics*.

Omsland, Anders, Diane C Cockrell, Dale Howe, Elizabeth R Fischer, Kimmo Virtaneva, Daniel E Sturdevant, Stephen F Porcella, and Robert A Heinzen (2009). "Host cell-free growth of the Q fever bacterium Coxiella burnetii." In: *Proceedings of the National Academy of Sciences* 106.11, pp. 4430–4434.

Overmann, Jörg (2005). "Chemotaxis and Behavioral Physiology of Not-Yet-Cultivated Microbes." In: *Methods in enzymology* 397, pp. 133–147.

Overmann, Jörg (2015). "Significance and future role of microbial resource centers." In: *Syst. Appl. Microbiol.* 38.4, pp. 258–265.

Overmann, Jörg, Birte Abt, and Johannes Sikorski (2017). "Present and Future of Culturing Bacteria." In: *Annual Review of Microbiology* 71, pp. 711–730.

Pace, Norman R. (2006). "Time for a change." In: *Nature* 441.7091, p. 289.

Pace, Norman R, David A Stahl, David J Lane, and Gary J Olsen (1986). "The analysis of natural microbial populations by ribosomal RNA sequences." In: *Advances in microbial ecology*. Springer, pp. 1–55.

Pagnier, Isabelle, Natalya Yutin, Olivier Croce, Kira S. Makarova, Yuri I. Wolf, Samia Benamar, Didier Raoult, Eugene V. Koonin, and Bernard La Scola (2015). "Babela massiliensis, a representative of a widespread bacterial phylum with unusual adaptations to parasitism in amoebae." In: *Biology Direct* 10.1, p. 13.

Paine, Robert T. (1966). "Food Web Complexity and Species Diversity." In: *The American Naturalist* 100.910, pp. 65–75.

Paliy, Oleg and Vijay Shankar (2016). "Application of multivariate statistical techniques in microbial ecology." In: *Molecular ecology* 25.5, pp. 1032–1057.

Park, Jihyang, Alissa Kerner, Mark A Burns, and Xiaoxia Nina Lin (2011). "Microdroplet-enabled highly parallel co-cultivation of microbial communities." In: *PloS one* 6.2, e17019.

Parker, Charles T, Brian J Tindall, and George M Garrity (2015). “International Code of Nomenclature of Prokaryotes.” In: *International journal of systematic and evolutionary microbiology*.

Parks, Donovan H., Maria Chuvochina, David W. Waite, Christian Rinke, Adam Skarshewski, Pierre-Alain Chaumeil, and Philip Hugenholtz (2018). “A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life.” In: *Nature Biotechnology* 36.10, pp. 996–1004.

Parniske, Martin (2008). “Arbuscular mycorrhiza: the mother of plant root endosymbioses.” In: *Nature Reviews. Microbiology* 6.10, pp. 763–775.

Parratt, Steven R and Anna-Liisa Laine (2016). “The role of hyperparasitism in microbial pathogen ecology and evolution.” In: *The ISME Journal* 10.8, pp. 1815–1822.

Pascal, Victoria, Marta Pozuelo, Natalia Borruel, Francesc Casellas, David Campos, Alba Santiago, Xavier Martinez, Encarna Varela, Guillaume Sarrabayrouse, Kathleen Machiels, Severine Vermeire, Harry Sokol, Francisco Guarner, and Chaysavanh Manichanh (2017). “A microbial signature for Crohn's disease.” In: *Gut* 66.5, pp. 813–822.

Pascual-García, Alberto, Javier Tamames, and Ugo Bastolla (2014). “Bacteria dialog with Santa Rosalia: Are aggregations of cosmopolitan bacteria mainly explained by habitat filtering or by ecological interactions?” In: *BMC microbiology* 14, p. 284.

Pasolli, Edoardo, Lucas Schiffer, Paolo Manghi, Audrey Renson, Valerie Obenchain, Duy Tin Truong, Francesco Beghini, Faizan Malik, Marcel Ramos, Jennifer B Dowd, Curtis Huttenhower, Martin Morgan, Nicola Segata, and Levi Waldron (2017). “Accessible, curated metagenomic data through ExperimentHub.” In: *Nat. Methods* 14.11, pp. 1023–1024.

Pasolli, Edoardo, Duy Tin Truong, Faizan Malik, Levi Waldron, and Nicola Segata (2016). “Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights.” In: *PLoS Comput. Biol.* 12.7, e1004977.

Paulson, Joseph N, O Colin Stine, Héctor Corrada Bravo, and Mihai Pop (2013). “Differential abundance analysis for microbial marker-gene surveys.” In: *Nature methods* 10.12, p. 1200.

Pawlowsky-Glahn, Vera, Juan José Egozcue, and Raimon Tolosana-Delgado (2015). *Modeling and Analysis of Compositional Data*. John Wiley & Sons.

Pearl, Judea (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference (Representation and Reasoning)*. San Mateo, CA: Morgan Kaufmann.

Pearl, Judea (2010). “An introduction to causal inference.” In: *The international journal of biostatistics* 6.2.

Pearson, Karl (1897). “Mathematical contributions to the theory of evolution.—On a form of spurious correlation which may arise when indices are used in the

measurement of organs." In: *Proceedings of the Royal Society of London* 60.359-367, pp. 489–498.

Peay, Kabir G., Peter G. Kennedy, and Thomas D. Bruns (2008). "Fungal Community Ecology: A Hybrid Beast with a Molecular Master." In: *BioScience* 58.9, pp. 799–810.

Pellissier, Loïc, Camille Albouy, Jordi Bascompte, Nina Farwig, Catherine Graham, Michel Loreau, Maria Alejandra Maglianesi, Carlos J Melián, Camille Pitteloud, Tomas Roslin, et al. (2018). "Comparing species interaction networks along environmental gradients." In: *Biological Reviews* 93.2, pp. 785–800.

Pérez-Cobas, Ana Elena, Andrés Moya, María José Gosalbes, and Amparo Latorre (2015). "Colonization Resistance of the Gut Microbiota against Clostridium difficile." In: *Antibiotics* 4.3, pp. 337–357.

Perna, Nicole T, Guy Plunkett III, Valerie Burland, Bob Mau, Jeremy D Glasner, Debra J Rose, George F Mayhew, Peter S Evans, Jason Gregor, Heather A Kirkpatrick, et al. (2001). "Genome sequence of enterohaemorrhagic Escherichia coli O157: H7." In: *Nature* 409.6819, p. 529.

Perras, Alexandra K., Gerhard Wanner, Andreas Klingl, Maximilian Mora, Anna K. Auerbach, Veronika Heinz, Alexander J. Probst, Harald Huber, Reinhard Rachel, Sandra Meck, and Christine Moissl-Eichinger (2014). "Grappling archaea: ultrastructural analyses of an uncultivated, cold-loving archaeon, and its biofilm." In: *Frontiers in Microbiology* 5, p. 397.

Peters, Brian M., Mary Ann Jabra-Rizk, Graeme A. O' May, J. William Costerton, and Mark E. Shirtliff (2012). "Polymicrobial Interactions: Impact on Pathogenesis and Human Disease." In: *Clinical Microbiology Reviews* 25.1, pp. 193–213.

Pierce, W Dwight, Robert Asa Cushman, Clifford E Hood, et al. (1912). "insect enemies of the cotton boll weevil." In:

Pimm, Stuart L. (1982). "Food webs." In: *Food webs*. Springer.

Pimm, Stuart L. (1984). "The complexity and stability of ecosystems." In: *Nature* 307.5949, pp. 321–326.

Pimm, Stuart L., John H. Lawton, and Joel E. Cohen (1991). "Food web patterns and their consequences." In: *Nature* 350.6320, pp. 669–674.

Pollet, Thomas, Lyria Berdjeb, Cédric Garnier, Gaël Durrieu, Christophe Le Poupon, Benjamin Misson, and Briand Jean-François (2018). "Prokaryotic community successions and interactions in marine biofilms: the key role of Flavobacteriia." In: *FEMS microbiology ecology* 94.6, fiy083.

Power, Mary E., David Tilman, James A. Estes, Bruce A. Menge, William J. Bond, L. Scott Mills, Gretchen Daily, Juan Carlos Castilla, Jane Lubchenco, and Robert T. Paine (1996). "Challenges in the Quest for Keystones." In: *BioScience* 46.8, pp. 609–620.

Props, Ruben, Frederiek-Maarten Kerckhof, Peter Rubbens, Jo De Vrieze, Emma Hernandez Sanabria, Willem Waegeman, Pieter Monsieurs, Frederik Hammes, and

Nico Boon (2017). “Absolute quantification of microbial taxon abundances.” In: *The ISME journal* 11.2, pp. 584–587.

Proulx, Stephen R., Daniel E. L. Promislow, and Patrick C. Phillips (2005). “Network thinking in ecology and evolution.” In: *Trends in Ecology & Evolution* 20.6, pp. 345–353.

Qin, Junjie, Yingrui Li, Zhiming Cai, Shenghui Li, Jianfeng Zhu, Fan Zhang, Suisha Liang, Wenwei Zhang, Yuanlin Guan, Dongqian Shen, Yangqing Peng, Dongya Zhang, Zhuye Jie, Wenxian Wu, Youwen Qin, Wenbin Xue, Junhua Li, Lingchuan Han, Donghui Lu, Peixian Wu, Yali Dai, Xiaojuan Sun, Zesong Li, Aifa Tang, Shilong Zhong, Xiaoping Li, Weineng Chen, Ran Xu, Mingbang Wang, Qiang Feng, Meihua Gong, Jing Yu, Yanyan Zhang, Ming Zhang, Torben Hansen, Gaston Sanchez, Jeroen Raes, Gwen Falony, Shujiro Okuda, Mathieu Almeida, Emmanuelle LeChatelier, Pierre Renault, Nicolas Pons, Jean-Michel Batto, Zhaoxi Zhang, Hua Chen, Ruifu Yang, Weimou Zheng, Songgang Li, Huanming Yang, Jian Wang, S Dusko Ehrlich, Rasmus Nielsen, Oluf Pedersen, Karsten Kristiansen, and Jun Wang (2012). “A metagenome-wide association study of gut microbiota in type 2 diabetes.” In: *Nature* 490.7418, pp. 55–60.

Quin, C, M Estaki, DM Vollman, JA Barnett, SK Gill, and DL Gibson (2018). “Probiotic supplementation and associated infant gut microbiome and health: a cautionary retrospective clinical comparison.” In: *Scientific reports* 8.1, p. 8283.

Quince, Christopher, Tom O Delmont, Sébastien Raguideau, Johannes Alneberg, Aaron E Darling, Gavin Collins, and A Murat Eren (2017). “DESMAN: a new tool for de novo extraction of strains from metagenomes.” In: *Genome biology* 18.1, p. 181.

Rabbani, Golam H., Shamsir Ahmed, Iqbal Hossain, Rafiqul Islam, Farzana Marni, Mastura Akhtar, and Nashiha Majid (2009). “Green banana reduces clinical severity of childhood shigellosis: a double-blind, randomized, controlled clinical trial.” In: *The Pediatric Infectious Disease Journal* 28.5, pp. 420–425.

Rampelotto, Pabulo Henrique (2013). “Extremophiles and Extreme Environments.” In: *Life : Open Access Journal* 3.3, pp. 482–485.

Rappé, Michael S., Stephanie A. Connon, Kevin L. Vergin, and Stephen J. Giovannoni (2002). “Cultivation of the ubiquitous SAR11 marine bacterioplankton clade.” In: *Nature* 418.6898, pp. 630–633.

Rawls, John F., Michael A. Mahowald, Ruth E. Ley, and Jeffrey I. Gordon (2006). “Reciprocal Gut Microbiota Transplants from Zebrafish and Mice to Germ-free Recipients Reveal Host Habitat Selection.” In: *Cell* 127.2, pp. 423–433.

Regier, Jeffrey, Kiran Pamnany, Keno Fischer, Andreas Noack, Maximilian Lam, Jarrett Revels, Steve Howard, Ryan Giordano, David Schlegel, Jon McAuliffe, et al. (2018). “Cataloging the Visible Universe through Bayesian Inference at Petascale.” In: *2018*

IEEE International Parallel and Distributed Processing Symposium (IPDPS). IEEE, pp. 44–53.

Ridenhour, Benjamin J, Sarah L Brooker, Janet E Williams, James T Van Leuven, Aaron W Miller, M Denise Dearing, and Christopher H Remien (2017). "Modeling time-series data from microbial communities." In: *ISME J.* 11.11, pp. 2526–2537.

Rinke, Christian, Patrick Schwientek, Alexander Sczyrba, Natalia N Ivanova, Iain J Anderson, Jan-Fang Cheng, Aaron Darling, Stephanie Malfatti, Brandon K Swan, Esther a Gies, Jeremy a Dodsworth, Brian P Hedlund, George Tsiamis, Stefan M Sievert, Wen-Tso Liu, Jonathan a Eisen, Steven J Hallam, Nikos C Kyrpides, Ramunas Stepanauskas, Edward M Rubin, Philip Hugenholtz, and Tanja Woyke (2013). "Insights into the phylogeny and coding potential of microbial dark matter." In: *Nature* 499.7459, pp. 431–7.

Robert, Christian P, Víctor Elvira, Nick Tawn, and Changye Wu (2018). "Accelerating MCMC algorithms." In: *Wiley Interdiscip. Rev. Comput. Stat.* 10.5, e1435.

Robinson, Mark D, Davis J McCarthy, and Gordon K Smyth (2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." In: *Bioinformatics* 26.1, pp. 139–140.

Rohr, Rudolf P. and Jordi Bascompte (2014). "Components of phylogenetic signal in antagonistic and mutualistic networks." In: *The American Naturalist* 184.5, pp. 556–564.

Rook, Graham A. W. (2007). "The hygiene hypothesis and the increasing prevalence of chronic inflammatory disorders." In: *Transactions of the Royal Society of Tropical Medicine and Hygiene* 101.11, pp. 1072–1074.

Rosindell, James, Stephen P. Hubbell, and Rampal S. Etienne (2011). "The unified neutral theory of biodiversity and biogeography at age ten." In: *Trends in Ecology & Evolution* 26.7, pp. 340–348.

Rosselló-Mora, Ramon and Rudolf Amann (2001). "The species concept for prokaryotes." In: *FEMS Microbiology Reviews* 25.1, pp. 39–67.

Röttjers, Lisa and Karoline Faust (2018). "From hairballs to hypotheses–biological insights from microbial networks." In: *FEMS Microbiology Reviews* 42.6, pp. 761–780.

Sacchi, Claudio T., Anne M. Whitney, Leonard W. Mayer, Roger Morey, Arnold Steigerwalt, Ariana Boras, Robin S. Weyant, and Tanja Popovic (2002). "Sequencing of 16S rRNA Gene: A Rapid Tool for Identification of Bacillus anthracis." In: *Emerging Infectious Diseases* 8.10, pp. 1117–1123.

Salonen, Anne, Janne Nikkilä, Jonna Jalanka-Tuovinen, Outi Immonen, Mirjana Rajilić-Stojanović, Riina A. Kekkonen, Airi Palva, and Willem M. de Vos (2010). "Comparative analysis of fecal DNA extraction methods with phylogenetic microarray:

effective recovery of bacterial and archaeal DNA using mechanical cell lysis." In: *Journal of Microbiological Methods* 81.2, pp. 127–134.

Salter, Susannah J., Michael J. Cox, Elena M. Turek, Szymon T. Calus, William O. Cookson, Miriam F. Moffatt, Paul Turner, Julian Parkhill, Nicholas J. Loman, and Alan W. Walker (2014). "Reagent and laboratory contamination can critically impact sequence-based microbiome analyses." In: *BMC Biology* 12.1, p. 87.

Samuel, Buck S., Abdullah Shaito, Toshiyuki Motoike, Federico E. Rey, Fredrik Backhed, Jill K. Manchester, Robert E. Hammer, S. Clay Williams, Jan Crowley, Masashi Yanagisawa, and Jeffrey I. Gordon (2008). "Effects of the gut microbiota on host adiposity are modulated by the short-chain fatty-acid binding G protein-coupled receptor, Gpr41." In: *Proceedings of the National Academy of Sciences of the United States of America* 105.43, pp. 16767–16772.

Sanders, Jon G., Annabel C. Beichman, Joe Roman, Jarrod J. Scott, David Emerson, James J. McCarthy, and Peter R. Girguis (2015). "Baleen whales host a unique gut microbiome with similarities to both carnivores and herbivores." In: *Nature Communications* 6, p. 8285.

Sanger, F., S. Nicklen, and A. R. Coulson (1977). "DNA sequencing with chain-terminating inhibitors." In: *Proceedings of the National Academy of Sciences of the United States of America* 74.12, pp. 5463–5467.

Sboner, Andrea and Constantin F. Aliferis (2005). "Modeling Clinical Judgment and Implicit Guideline Compliance in the Diagnosis of Melanomas Using Machine Learning." In: *AMIA Annual Symposium Proceedings* 2005, pp. 664–668.

Scherlach, Kirstin and Christian Hertweck (2018). "Mediators of mutualistic microbe–microbe interactions." In: *Natural Product Reports* 35.4, pp. 303–308.

Schloss, Patrick D (2010). "The effects of alignment quality, distance calculation method, sequence filtering, and region on the analysis of 16S rRNA gene-based studies." In: *PLoS Comput. Biol.* 6.7, e1000844.

Schloss, Patrick D., Rene A. Girard, Thomas Martin, Joshua Edwards, and J. Cameron Thrash (2016). "Status of the Archaeal and Bacterial Census: an Update." In: *mBio* 7.3, e00201–16.

Schloss, Patrick D and Sarah L Westcott (2011). "Assessing and improving methods used in operational taxonomic unit-based approaches for 16S rRNA gene sequence analysis." In: *Appl. Environ. Microbiol.* 77.10, pp. 3219–3226.

Schloss, Patrick D, Sarah L Westcott, Thomas Ryabin, Justine R Hall, Martin Hartmann, Emily B Hollister, Ryan A Lesniewski, Brian B Oakley, Donovan H Parks, Courtney J Robinson, Jason W Sahl, Blaz Stres, Gerhard G Thallinger, David J Van Horn, and Carolyn F Weber (2009). "Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities." In: *Appl. Environ. Microbiol.* 75.23, pp. 7537–7541.

Schlüter, Federico (2012). “A survey on independence-based Markov networks learning.” In: *Artificial Intelligence Review* 42.4, pp. 1069–1093.

Schmidt, Thomas S B, João F Matias Rodrigues, and Christian von Mering (2014). “Ecological Consistency of SSU rRNA-Based Operational Taxonomic Units at a Global Scale.” In: *PLoS Comput. Biol.* 10.4, e1003594.

Schmidt, Thomas S B, João F Matias Rodrigues, and Christian von Mering (2015). “Limits to robustness and reproducibility in the demarcation of operational taxonomic units.” In: *Environ. Microbiol.* 17.5, pp. 1689–1706.

Schmidt, Thomas S. B., Jeroen Raes, and Peer Bork (2018). “The Human Gut Microbiome: From Association to Modulation.” In: *Cell* 172.6, pp. 1198–1215.

Schmidt, Thomas Sebastian Benedikt, João Frederico Matias Rodrigues, and Christian von Mering (2017). “A family of interaction-adjusted indices of community similarity.” In: *ISME J.* 11.3, pp. 791–807.

Schmitt, CL and ML Tatum (2008). “The Malheur National Forest Location of the World’s Largest Living Organism [The Humongous Fungus].” In: *Agriculture Forest Service Pacific Northwest Region* 17.11, p. 2015.

Schwager, Emma, Himel Mallick, Steffen Ventz, and Curtis Huttenhower (2017). “A Bayesian method for detecting pairwise associations in compositional data.” In: *PLoS Comput. Biol.* 13.11, e1005852.

Sczyrba, Alexander, Peter Hofmann, Peter Belmann, David Koslicki, Stefan Janssen, Johannes Dröge, Ivan Gregor, Stephan Majda, Jessika Fiedler, Eik Dahms, Andreas Bremges, Adrian Fritz, Ruben Garrido-Oter, Tue Sparholt Jørgensen, Nicole Shapiro, Philip D Blood, Alexey Gurevich, Yang Bai, Dmitrij Turaev, Matthew Z DeMaere, Rayan Chikhi, Niranjan Nagarajan, Christopher Quince, Fernando Meyer, Monika Balvo\ v ciūtė, Lars Hestbjerg Hansen, Søren J Sørensen, Burton K H Chia, Bertrand Denis, Jeff L Froula, Zhong Wang, Robert Egan, Dongwan Don Kang, Jeffrey J Cook, Charles Deltel, Michael Beckstette, Claire Lemaitre, Pierre Peterlongo, Guillaume Rizk, Dominique Lavenier, Yu-Wei Wu, Steven W Singer, Chirag Jain, Marc Strous, Heiner Klingenberg, Peter Meinicke, Michael D Barton, Thomas Lingner, Hsin-Hung Lin, Yu-Chieh Liao, Genivaldo Gueiros Z Silva, Daniel A Cuevas, Robert A Edwards, Surya Saha, Vitor C Piro, Bernhard Y Renard, Mihai Pop, Hans-Peter Klenk, Markus Göker, Nikos C Kyrpides, Tanja Woyke, Julia A Vorholt, Paul Schulze-Lefert, Edward M Rubin, Aaron E Darling, Thomas Rattei, and Alice C McHardy (2017). “Critical Assessment of Metagenome Interpretation-a benchmark of metagenomics software.” In: *Nat. Methods* 14.11, pp. 1063–1071.

Segata, Nicola (2018). “On the road to strain-resolved comparative metagenomics.” In: *mSystems* 3.2, e00190–17.

Segata, Nicola, Jacques Izard, Levi Waldron, Dirk Gevers, Larisa Miropolsky, Wendy S Garrett, and Curtis Huttenhower (2011). "Metagenomic biomarker discovery and explanation." In: *Genome biology* 12.6, R60.

Seng, Piseth, Cedric Abat, Jean Marc Rolain, Philippe Colson, Jean-Christophe Lagier, Frédérique Gouriet, Pierre Edouard Fournier, Michel Drancourt, Bernard La Scola, and Didier Raoult (2013). "Identification of Rare Pathogenic Bacteria in a Clinical Microbiology Laboratory: Impact of Matrix-Assisted Laser Desorption Ionization–Time of Flight Mass Spectrometry." In: *Journal of Clinical Microbiology* 51.7, pp. 2182–2194.

Shade, Ashley, Hannes Peter, Steven D Allison, Didier Baho, Mercè Berga, Helmut Bürgmann, David H Huber, Silke Langenheder, Jay T Lennon, Jennifer BH Martiny, et al. (2012). "Fundamentals of microbial community resistance and resilience." In: *Frontiers in microbiology* 3, p. 417.

Shapiro, B. Jesse, Jean-Baptiste Leducq, and James Mallet (2016). "What Is Speciation?" In: *PLOS Genetics* 12.3, e1005860.

Shapiro, B. Jesse and Martin F. Polz (2014). "Ordering microbial diversity into ecologically and genetically cohesive units." In: *Trends in Microbiology* 22.5, pp. 235–247.

Shapiro, J. A. (1998). "Thinking about bacterial populations as multicellular organisms." In: *Annual Review of Microbiology* 52, pp. 81–104.

Sharma, Sanjib (2017). "Markov Chain Monte Carlo Methods for Bayesian Data Analysis in Astronomy." In: *Annu. Rev. Astron. Astrophys.* 55.1, pp. 213–259.

Sharon, Itai, Michael Kertesz, Laura A. Hug, Dmitry Pushkarev, Timothy A. Blauwkamp, Cindy J. Castelle, Mojgan Amirebrahimi, Brian C. Thomas, David Burstein, Susannah G. Tringe, Kenneth H. Williams, and Jillian F. Banfield (2015). "Accurate, multi-kb reads resolve complex populations and detect rare microorganisms." In: *Genome Research* 25.4, pp. 534–543.

Sharpton, Thomas J (2018). "Role of the gut microbiome in vertebrate evolution." In: *MSystems* 3.2, e00174–17.

Shaw, Grace Tzun-Wen, Yueh-Yang Pao, and Daryi Wang (2016). "MetaMIS: a metagenomic microbial interaction simulator based on microbial community profiles." In: *BMC Bioinformatics* 17.1, p. 488.

Sinha, Rashmi, Galeb Abu-Ali, Emily Vogtmann, Anthony A Fodor, Boyu Ren, Amnon Amir, Emma Schwager, Jonathan Crabtree, Siyuan Ma, Microbiome Quality Control Project Consortium, Christian C Abnet, Rob Knight, Owen White, and Curtis Huttenhower (2017). "Assessment of variation in microbial community amplicon sequencing by the Microbiome Quality Control (MBQC) project consortium." In: *Nat. Biotechnol.* 35.11, pp. 1077–1086.

Slavin, Joanne and Joanne Slavin (2013). “Fiber and Prebiotics: Mechanisms and Health Benefits.” In: *Nutrients* 5.4, pp. 1417–1435.

Sloan, William T., Mary Lunn, Stephen Woodcock, Ian M. Head, Sean Nee, and Thomas P. Curtis (2006). “Quantifying the roles of immigration and chance in shaping prokaryote community structure.” In: *Environmental Microbiology* 8.4, pp. 732–740.

Sogin, Mitchell L, Hilary G Morrison, Julie A Huber, David Mark Welch, Susan M Huse, Phillip R Neal, Jesus M Arrieta, and Gerhard J Herndl (2006). “Microbial diversity in the deep sea and the underexplored “rare biosphere”.” In: *Proc. Natl. Acad. Sci. U. S. A.* 103.32, pp. 12115–12120.

Solé, Ricard V. and M. Montoya (2001). “Complexity and fragility in ecological networks.” In: *Proceedings of the Royal Society of London B: Biological Sciences* 268.1480, pp. 2039–2045.

Solé, Ricard V and Josep Sardanyés (2014). “Red Queen Coevolution on Fitness Landscapes.” In: *Emergence, Complexity and Computation*, pp. 301–338.

Sonnenburg, Justin L., Christina T. L. Chen, and Jeffrey I. Gordon (2006). “Genomic and Metabolic Studies of the Impact of Probiotics on a Model Gut Symbiont and Host.” In: *PLOS Biology* 4.12, e413.

Soucy, Shannon M., Jinling Huang, and Johann Peter Gogarten (2015). “Horizontal gene transfer: building the web of life.” In: *Nature Reviews Genetics* 16.8, pp. 472–482.

Spirtes, Peter (2010). “Introduction to causal inference.” In: *Journal of Machine Learning Research* 11.5, pp. 1643–1662.

Spirtes, Peter, Clark N Glymour, Richard Scheines, David Heckerman, Christopher Meek, Gregory Cooper, and Thomas Richardson (2000). *Causation, prediction, and search*. MIT press.

Stackebrandt, Erko (2006). “Taxonomic parameters revisited: tarnished gold standards.” In: *Microbiol. Today* 33, pp. 152–155.

Stahl, David A, David J Lane, Gary J Olsen, and Norman R Pace (1985). “Characterization of a Yellowstone hot spring microbial community by 5S rRNA sequences.” In: *Appl. Environ. Microbiol.* 49.6, pp. 1379–1384.

Staley, James T and Allan Konopka (1985). “Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats.” In: *Annual review of microbiology* 39.1, pp. 321–346.

Stanley, Claire E. and Marcel G. A. van der Heijden (2017). “Microbiome-on-a-Chip: New Frontiers in Plant–Microbiota Research.” In: *Trends in Microbiology* 25.8, pp. 610–613.

Stat, Michael, Emily Morris, and Ruth D. Gates (2008). “Functional diversity in coral–dinoflagellate symbiosis.” In: *Proceedings of the National Academy of Sciences* 105.27, pp. 9256–9261.

Statnikov, Alexander, Alexander V Alekseyenko, Zhiguo Li, Mikael Henaff, Guillermo I Perez-Perez, Martin J Blaser, and Constantin F Aliferis (2013a). "Microbiomic signatures of psoriasis: feasibility and methodology comparison." In: *Sci. Rep.* 3, p. 2620.

Statnikov, Alexander, Mikael Henaff, Varun Narendra, Kranti Konganti, Zhiguo Li, Liying Yang, Zhiheng Pei, Martin J Blaser, Constantin F Aliferis, and Alexander V Alekseyenko (2013b). "A comprehensive evaluation of multicategory classification methods for microbiomic data." In: *Microbiome* 1.1, p. 11.

Stecher, Bärbel, Samuel Chaffron, Rina Käppeli, Siegfried Hapfelmeier, Susanne Freedrich, Thomas C. Weber, Jorum Kirundi, Mrutyunjay Suar, Kathy D. McCoy, Christian von Mering, Andrew J. Macpherson, and Wolf-Dietrich Hardt (2010). "Like Will to Like: Abundances of Closely Related Species Can Predict Susceptibility to Intestinal Colonization by Pathogenic and Commensal Bacteria." In: *PLOS Pathogens* 6.1, e1000711.

Stein, Richard R., Vanni Bucci, Nora C. Toussaint, Charlie G. Buffie, Gunnar Rätsch, Eric G. Pamer, Chris Sander, and João B. Xavier (2013). "Ecological Modeling from Time-Series Inference: Insight into Dynamics and Stability of Intestinal Microbiota." In: *PLOS Computational Biology* 9.12, e1003388.

Stepanauskas, Ramunas (2012). "Single cell genomics: an individual look at microbes." In: *Current opinion in microbiology* 15.5, pp. 613–20.

Stolovitzky, Gustavo, Don Monroe, and Andrea Califano (2007). "Dialogue on reverse-engineering assessment and methods: the DREAM of high-throughput pathway inference." In: *Ann. N. Y. Acad. Sci.* 1115, pp. 1–22.

Strassmann, Joan E., Owen M. Gilbert, and David C. Queller (2011). "Kin discrimination and cooperation in microbes." In: *Annual Review of Microbiology* 65, pp. 349–367.

Strogatz, S. H. (2001). "Exploring complex networks." In: *Nature* 410.6825, pp. 268–276.

Suez, Jotham, Niv Zmora, Gili Zilberman-Schapira, Uria Mor, Mally Dori-Bachash, Stavros Bashiardes, Maya Zur, Dana Regev-Lehavi, Rotem Ben-Zeev Brik, Sara Federici, Max Horn, Yotam Cohen, Andreas E. Moor, David Zeevi, Tal Korem, Eran Kotler, Alon Harmelin, Shalev Itzkovitz, Nitsan Maharshak, Oren Shibolet, Meirav Pevsner-Fischer, Hagit Shapiro, Itai Sharon, Zamir Halpern, Eran Segal, and Eran Elinav (2018). "Post-Antibiotic Gut Mucosal Microbiome Reconstitution Is Impaired by Probiotics and Improved by Autologous FMT." In: *Cell* 174.6, 1406–1423.e16.

Sullivan, Matthew B, John B Waterbury, and Sallie W Chisholm (2003). "Cyanophages infecting the oceanic cyanobacterium Prochlorococcus." In: *Nature* 424.6952, pp. 1047–1051.

Sun, Dong-Lei, Xuan Jiang, Qinglong L. Wu, and Ning-Yi Zhou (2013). “Intragenomic heterogeneity of 16S rRNA genes causes overestimation of prokaryotic diversity.” In: *Applied and Environmental Microbiology* 79.19, pp. 5962–5969.

Sunagawa, Shinichi, Luis Pedro Coelho, Samuel Chaffron, Jens Roat Kultima, Karine Labadie, Guillem Salazar, Bardya Djahanschiri, Georg Zeller, Daniel R Mende, Adriana Alberti, et al. (2015). “Structure and function of the global ocean microbiome.” In: *Science* 348.6237, p. 1261359.

Sunagawa, Shinichi, Daniel R Mende, Georg Zeller, Fernando Izquierdo-Carrasco, Simon A Berger, Jens Roat Kultima, Luis Pedro Coelho, Manimozhiyan Arumugam, Julien Tap, Henrik Bjørn Nielsen, et al. (2013). “Metagenomic species profiling using universal phylogenetic marker genes.” In: *Nature methods* 10.12, p. 1196.

Szewzyk, Ulrich, Regine Szewzyk, and TA Stenström (1994). “Thermophilic, anaerobic bacteria isolated from a deep borehole in granite in Sweden.” In: *Proceedings of the national academy of sciences* 91.5, pp. 1810–1813.

Tackmann, Janko, Natasha Arora, T. S. B. Schmidt, João F. M. Rodrigues, and Christian von Mering (2018). “Ecologically informed microbial biomarkers and accurate classification of mixed and unmixed samples in an extensive cross-study of human body sites.” In: *Microbiome* 6.1, p. 192.

Tang, Lei (2019). “Microbial interactions.” In: *Nature methods* 16.1, p. 19.

Tankou, Stephanie K., Keren Regev, Brian C. Healy, Emily Tjon, Luca Laghi, Laura M. Cox, Pia Kivisäkk, Isabelle V. Pierre, Lokhande Hrishikesh, Roopali Gandhi, Sandra Cook, Bonnie Glanz, James Stankiewicz, and Howard L. Weiner (2018). “A probiotic modulates the microbiome and immunity in multiple sclerosis.” In: *Annals of Neurology* 83.6, pp. 1147–1161.

Thompson, John N (1994). *The coevolutionary process*. University of Chicago Press.

Thompson, John N (2005). *The geographic mosaic of coevolution*. University of Chicago Press.

Thompson, Luke R, Jon G Sanders, Daniel McDonald, Amnon Amir, Joshua Ladau, Kenneth J Locey, Robert J Prill, Anupriya Tripathi, Sean M Gibbons, Gail Ackermann, Jose A Navas-Molina, Stefan Janssen, Evguenia Kopylova, Yoshiki Vázquez-Baeza, Antonio González, James T Morton, Siavash Mirarab, Zhenjiang Zech Xu, Lingjing Jiang, Mohamed F Haroon, Jad Kanbar, Qiyun Zhu, Se Jin Song, Tomasz Kosciolek, Nicholas A Bokulich, Joshua Lefler, Colin J Brislawn, Gregory Humphrey, Sarah M Owens, Jarrad Hampton-Marcell, Donna Berg-Lyons, Valerie McKenzie, Noah Fierer, Jed A Fuhrman, Aaron Clauset, Rick L Stevens, Ashley Shade, Katherine S Pollard, Kelly D Goodwin, Janet K Jansson, Jack A Gilbert, Rob Knight, and Earth Microbiome Project Consortium (2017). “A communal catalogue reveals Earth's multiscale microbial diversity.” In: *Nature* 551.7681, pp. 457–463.

Tirumalai, Madhan R., Victor G. Stepanov, Andrea Wünsche, Saied Montazari, Racquel O. Gonzalez, Kasturi Venkateswaran, and George E. Fox (2018). "Bacillus safensis FO-36b and Bacillus pumilus SAFR-032: a whole genome comparison of two spacecraft assembly facility isolates." In: *BMC Microbiology* 18.1, p. 57.

Tracanna, Vittorio, Anne de Jong, Marnix H. Medema, and Oscar P. Kuipers (2017). "Mining prokaryotes for antimicrobial compounds: from diversity to function." In: *FEMS Microbiology Reviews* 41.3, pp. 417–429.

Tremblay, Julien, Kanwar Singh, Alison Fern, Edward S. Kirton, Shaomei He, Tanja Woyke, Janey Lee, Feng Chen, Jeffery L. Dangl, and Susannah G. Tringe (2015). "Primer and platform effects on 16S rRNA tag sequencing." In: *Frontiers in Microbiology* 6.

Treseder, Kathleen K., Teri C. Balser, Mark A. Bradford, Eoin L. Brodie, Eric A. Dubinsky, Valerie T. Eviner, Kirsten S. Hofmockel, Jay T. Lennon, Uri Y. Levine, Barbara J. MacGregor, Jennifer Pett-Ridge, and Mark P. Waldrop (2012). "Integrating microbial ecology into ecosystem models: challenges and priorities." In: *Biogeochemistry* 109.1, pp. 7–18.

Tripp, H. James, Joshua B. Kitner, Michael S. Schwalbach, John W. H. Dacey, Larry J. Wilhelm, and Stephen J. Giovannoni (2008). "SAR11 marine bacteria require exogenous reduced sulphur for growth." In: *Nature* 452.7188, pp. 741–744.

Troyer, Katherine (1984). "Microbes, herbivory and the evolution of social behavior." In: *Journal of Theoretical Biology* 106.2, pp. 157–169.

Truong, Duy Tin, Eric A. Franzosa, Timothy L. Tickle, Matthias Scholz, George Weingart, Edoardo Pasolli, Adrian Tett, Curtis Huttenhower, and Nicola Segata (2015). "MetaPhlAn2 for enhanced metagenomic taxonomic profiling." In: *Nature Methods* 12.10, pp. 902–903.

Truong, Duy Tin, Adrian Tett, Edoardo Pasolli, Curtis Huttenhower, and Nicola Segata (2017). "Microbial strain-level population structure and genetic diversity from metagenomes." In: *Genome Res.* 27.4, pp. 626–638.

Tsai, Chung-Jung, Buyong Ma, and Ruth Nussinov (2009). "Protein–protein interaction networks: how can a hub protein bind so many different partners?" In: *Trends Biochem. Sci.* 34.12, pp. 594–600.

Tsamardinos, Ioannis, Constantin F. Aliferis, and Alexander Statnikov (2003a). "Time and Sample Efficient Discovery of Markov Blankets and Direct Causal Relations." In: *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '03. New York, NY, USA: ACM, pp. 673–678.

Tsamardinos, Ioannis, Constantin F Aliferis, Alexander R Statnikov, and Er Statnikov (2003b). "Algorithms for Large Scale Markov Blanket Discovery." In: *FLAIRS conference*. Vol. 2, pp. 376–380.

Tsamardinos, Ioannis, Laura E Brown, and Constantin F Aliferis (2006). “The max-min hill-climbing Bayesian network structure learning algorithm.” In: *Mach. Learn.* 65.1, pp. 31–78.

Tsilimigras, Matthew C B and Anthony A Fodor (2016). “Compositional data analysis of the microbiome: fundamentals, tools, and challenges.” In: *Ann. Epidemiol.* 26.5, pp. 330–335.

Turnbaugh, Peter J, Micah Hamady, Tanya Yatsunenko, Brandi L Cantarel, Alexis Duncan, Ruth E Ley, Mitchell L Sogin, William J Jones, Bruce A Roe, Jason P Affourtit, et al. (2009). “A core gut microbiome in obese and lean twins.” In: *nature* 457.7228, p. 480.

Vaishampayan, Parag A., Elke Rabbow, Gerda Horneck, and Kasthuri J. Venkateswaran (2012). “Survival of Bacillus pumilus spores for a prolonged period of time in real space conditions.” In: *Astrobiology* 12.5, pp. 487–497.

Vaishnava, Shipra, Cassie L. Behrendt, Anisa S. Ismail, Lars Eckmann, and Lora V. Hooper (2008). “Paneth cells directly sense gut commensals and maintain homeostasis at the intestinal host-microbial interface.” In: *Proceedings of the National Academy of Sciences of the United States of America* 105.52, pp. 20858–20863.

Vajda, S, Claude E Shannon, and Warren Weaver (1950). “The Mathematical Theory of Communication.” In: *The Mathematical Gazette* 34.310, p. 312.

Valm, Alex M., Jessica L. Mark Welch, Christopher W. Rieken, Yuko Hasegawa, Mitchell L. Sogin, Rudolf Oldenbourg, Floyd E. Dewhirst, and Gary G. Borisy (2011). “Systems-level analysis of microbial community organization through combinatorial labeling and spectral imaging.” In: *Proceedings of the National Academy of Sciences of the United States of America* 108.10, pp. 4152–4157.

Vandeputte, Doris, Gunter Kathagen, Kevin D' hoe, Sara Vieira-Silva, Mireia Valles-Colomer, João Sabino, Jun Wang, Raul Y Tito, Lindsey De Commer, Youssef Darzi, Séverine Vermeire, Gwen Falony, and Jeroen Raes (2017a). “Quantitative microbiome profiling links gut community variation to microbial load.” In: *Nature* 551.7681, pp. 507–511.

Vandeputte, Doris, Raul Y. Tito, Rianne Vanleeuwen, Gwen Falony, and Jeroen Raes (2017b). “Practical considerations for large-scale gut microbiome studies.” In: *FEMS Microbiology Reviews* 41.Supp_1, S154–S167.

Varghese, Neha J, Supratim Mukherjee, Natalia Ivanova, Konstantinos T Konstantinidis, Kostas Mavrommatis, Nikos C Kyrpides, and Amrita Pati (2015). “Microbial species delineation using whole genome sequences.” In: *Nucleic Acids Res.* 43.14, pp. 6761–6771.

Venter, J. Craig, Karin Remington, John F. Heidelberg, Aaron L. Halpern, Doug Rusch, Jonathan A. Eisen, Dongying Wu, Ian Paulsen, Karen E. Nelson, William Nelson, Derrick E. Fouts, Samuel Levy, Anthony H. Knap, Michael W. Lomas, Ken Nealson,

Owen White, Jeremy Peterson, Jeff Hoffman, Rachel Parsons, Holly Baden-Tillson, Cynthia Pfannkoch, Yu-Hui Rogers, and Hamilton O. Smith (2004). "Environmental Genome Shotgun Sequencing of the Sargasso Sea." In: *Science* 304.5667, pp. 66–74.

Vermeesch, Pieter (2006). "Tectonic discrimination diagrams revisited." In: *Geochemistry, Geophysics, Geosystems* 7.6.

Větrovskỳ, Tomáš and Petr Baldrian (2013). "The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses." In: *PloS one* 8.2, e57923.

Vilcinskas, Andreas, Kilian Stoecker, Henrike Schmidtberg, Christian R. Röhrich, and Heiko Vogel (2013). "Invasive Harlequin Ladybird Carries Biological Weapons Against Native Competitors." In: *Science* 340.6134, pp. 862–863.

Vorholt, Julia A., Christine Vogel, Charlotte I. Carlström, and Daniel B. Müller (2017). "Establishing Causality: Opportunities of Synthetic Communities for Plant Microbiome Research." In: *Cell Host & Microbe* 22.2, pp. 142–155.

Wadsworth, W Duncan, Raffaele Argiento, Michele Guindani, Jessica Galloway-Pena, Samuel A Shelburne, and Marina Vannucci (2017). "An integrative Bayesian Dirichlet-multinomial regression model for the analysis of taxonomic abundances in microbiome data." In: *BMC bioinformatics* 18.1, p. 94.

Walker, Alan W., Jennifer C. Martin, Paul Scott, Julian Parkhill, Harry J. Flint, and Karen P. Scott (2015). "16S rRNA gene-based profiling of the human infant gut microbiota is strongly influenced by sample processing and PCR primer choice." In: *Microbiome* 3, p. 26.

Walker, JCG (1980). "The oxygen cycle in the natural environment and the biogeochemical cycles." In: *Berlin, Federal Republic of Germany: Springer*.

Walters, William, Embriette R Hyde, Donna Berg-Lyons, Gail Ackermann, Greg Humphrey, Alma Parada, Jack A Gilbert, Janet K Jansson, J Gregory Caporaso, Jed A Fuhrman, et al. (2016). "Improved bacterial 16S rRNA gene (V4 and V4-5) and fungal internal transcribed spacer marker gene primers for microbial community surveys." In: *Msystems* 1.1, e00009–15.

Wang, Jun and Huijue Jia (2016). "Metagenome-wide association studies: fine-mining the microbiome." In: *Nat. Rev. Microbiol.* 14.8, pp. 508–522.

Wang, Qiong, George M Garrity, James M Tiedje, and James R Cole (2007). "Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy." In: *Appl. Environ. Microbiol.* 73.16, pp. 5261–5267.

Ward, David M, Roland Weller, and Mary M Bateson (1990). "16S rRNA sequences reveal numerous uncultured microorganisms in a natural community." In: *Nature* 345.6270, p. 63.

Ward, Doyle V, Matthias Scholz, Moreno Zolfo, Diana H Taft, Kurt R Schibler, Adrian Tett, Nicola Segata, and Ardythe L Morrow (2016). "Metagenomic Sequencing with

Strain-Level Resolution Implicates Uropathogenic E. coli in Necrotizing Enterocolitis and Mortality in Preterm Infants." In: *Cell Rep.* 14.12, pp. 2912–2924.

Waters, Christopher M. and Bonnie L. Bassler (2005). "Quorum sensing: cell-to-cell communication in bacteria." In: *Annual Review of Cell and Developmental Biology* 21, pp. 319–346.

Waters, Elizabeth, Michael J. Hohn, Ivan Ahel, David E. Graham, Mark D. Adams, Mary Barnstead, Karen Y. Beeson, Lisa Bibbs, Randall Bolanos, Martin Keller, Keith Kretz, Xiaoying Lin, Eric Mathur, Jingwei Ni, Mircea Podar, Toby Richardson, Granger G. Sutton, Melvin Simon, Dieter Söll, Karl O. Stetter, Jay M. Short, and Michiel Noordewier (2003). "The genome of Nanoarchaeum equitans: Insights into early archaeal evolution and derived parasitism." In: *Proceedings of the National Academy of Sciences* 100.22, pp. 12984–12988.

Watson, James D, Francis HC Crick, et al. (1953). "Molecular structure of nucleic acids." In: *Nature* 171.4356, pp. 737–738.

Wattam, Alice R, James J Davis, Rida Assaf, Sébastien Boisvert, Thomas Brettin, Christopher Bun, Neal Conrad, Emily M Dietrich, Terry Disz, Joseph L Gabbard, Svetlana Gerdes, Christopher S Henry, Ronald W Kenyon, Dustin Machi, Chunhong Mao, Eric K Nordberg, Gary J Olsen, Daniel E Murphy-Olson, Robert Olson, Ross Overbeek, Bruce Parrello, Gordon D Pusch, Maulik Shukla, Veronika Vonstein, Andrew Warren, Fangfang Xia, Hyunseung Yoo, and Rick L Stevens (2017). "Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center." In: *Nucleic Acids Res.* 45.D1, pp. D535–D542.

Wayne, LG, DJ Brenner, RR Colwell, PAD Grimont, O Kandler, MI Krichevsky, LH Moore, WEC Moore, RGEea Murray, ESMP Stackebrandt, et al. (1987). "Report of the ad hoc committee on reconciliation of approaches to bacterial systematics." In: *International Journal of Systematic and Evolutionary Microbiology* 37.4, pp. 463–464.

Weiss, Sophie, Will Van Treuren, Catherine Lozupone, Karoline Faust, Jonathan Friedman, Ye Deng, Li Charlie Xia, Zhenjiang Zech Xu, Luke Ursell, Eric J Alm, Amanda Birmingham, Jacob A Cram, Jed A Fuhrman, Jeroen Raes, Fengzhu Sun, Jizhong Zhou, and Rob Knight (2016). "Correlation detection strategies in microbial data sets vary widely in sensitivity and precision." In: *ISME J.* 10.7, pp. 1669–1681.

Weiss, Sophie, Zhenjiang Zech Xu, Shyamal Peddada, Amnon Amir, Kyle Bittinger, Antonio Gonzalez, Catherine Lozupone, Jesse R Zaneveld, Yoshiki Vázquez-Baeza, Amanda Birmingham, Embriette R Hyde, and Rob Knight (2017). "Normalization and microbial differential abundance strategies depend upon data characteristics." In: *Microbiome* 5.1, p. 27.

West, Stuart A. and Angus Buckling (2003). "Cooperation, virulence and siderophore production in bacterial parasites." In: *Proceedings of the Royal Society B: Biological Sciences* 270.1510, pp. 37–44.

West, Stuart A., Ashleigh S. Griffin, Andy Gardner, and Stephen P. Diggle (2006). "Social evolution theory for microorganisms." In: *Nature Reviews Microbiology* 4.8, pp. 597–607.

Westcott, Sarah L and Patrick D Schloss (2015). "De novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units." In: *PeerJ* 3, e1487.

Whiteley, Marvin, Kimberly M Lee, and EP Greenberg (1999). "Identification of genes controlled by quorum sensing in Pseudomonas aeruginosa." In: *Proceedings of the National Academy of Sciences* 96.24, pp. 13904–13909.

Whittaker, Robert H (1972). "Evolution and measurement of species diversity." In: *Taxon*, pp. 213–251.

Widder, Edith A (2010). "Bioluminescence in the ocean: origins of biological, chemical, and ecological diversity." In: *Science* 328.5979, pp. 704–708.

Wikimedia Commons (2005). *File:Legionella Plate 01.png — Wikimedia Commons, the free media repository. Accessed 30-January-2019*. `https://upload.wikimedia.org/wikipedia/commons/9/93/Legionella_Plate_01.png`.

Wikimedia Commons (2006). *File:SimpleBayesNet.svg — Wikimedia Commons, the free media repository. Accessed 30-January-2019*. `https://upload.wikimedia.org/wikipedia/commons/0/0e/SimpleBayesNet.svg`.

Wikimedia Commons (2009). *File:Nitrogen_Cycle.svg — Wikimedia Commons, the free media repository. Accessed 30-January-2019*. `https://upload.wikimedia.org/wikipedia/commons/f/fe/Nitrogen_Cycle.svg`.

Wilhelm, Roland C., Kristin J. Radtke, Nadia C. S. Mykytczuk, Charles W. Greer, and Lyle G. Whyte (2012). "Life at the wedge: the activity and diversity of arctic ice wedge microbial communities." In: *Astrobiology* 12.4, pp. 347–360.

Williams, Richard J., Eric L. Berlow, Jennifer A. Dunne, Albert-László Barabási, and Neo D. Martinez (2002). "Two degrees of separation in complex food webs." In: *Proceedings of the National Academy of Sciences* 99.20, pp. 12913–12916.

Williams, RJP (1997). "The natural selection of the chemical elements." In: *Cellular and Molecular Life Sciences CMLS* 53.10, pp. 816–829.

Willing, Ben, Jonas Halfvarson, Johan Dicksved, Magnus Rosenquist, Gunnar Järnerot, Lars Engstrand, Curt Tysk, and Janet K. Jansson (2009). "Twin studies reveal specific imbalances in the mucosa-associated microbiota of patients with ileal Crohn's disease." In: *Inflammatory Bowel Diseases* 15.5, pp. 653–660.

Wingreen, Ned S and Simon A Levin (2006). "Cooperation among microorganisms." In: *PLoS biology* 4.9, e299.

Woese, Carl R (2002). "On the evolution of cells." In: *Proceedings of the National Academy of Sciences* 99.13, pp. 8742–8747.

Woese, Carl R and George E Fox (1977). “Phylogenetic structure of the prokaryotic domain: the primary kingdoms.” In: *Proceedings of the National Academy of Sciences* 74.11, pp. 5088–5090.

Woese, Carl R, Otto Kandler, and Mark L Wheelis (1990). “Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya.” In: *Proceedings of the National Academy of Sciences* 87.12, pp. 4576–4579.

Woodcock, Stephen, Christopher J. van der Gast, Thomas Bell, Mary Lunn, Thomas P. Curtis, Ian M. Head, and William T. Sloan (2007). “Neutral assembly of bacterial communities.” In: *FEMS microbiology ecology* 62.2, pp. 171–180.

Woodward, Guy, Bo Ebenman, Mark Emmerson, Jose M Montoya, Jens M Olesen, Alfredo Valido, and Philip H Warren (2005). “Body size in ecological networks.” In: *Trends in ecology & evolution* 20.7, pp. 402–409.

Wootton, J. Timothy (2005). “Field parameterization and experimental test of the neutral theory of biodiversity.” In: *Nature* 433.7023, pp. 309–312.

Worden, Alexandra Z, Michael J Follows, Stephen J Giovannoni, Susanne Wilken, Amy E Zimmerman, and Patrick J Keeling (2015). “Rethinking the marine carbon cycle: factoring in the multifarious lifestyles of microbes.” In: *Science* 347.6223, p. 1257594.

Xia, Li C, Joshua A Steele, Jacob A Cram, Zoe G Cardon, Sheri L Simmons, Joseph J Vallino, Jed A Fuhrman, and Fengzhu Sun (2011). “Extended local similarity analysis (eLSA) of microbial community and other time series data with replicates.” In: *BMC systems biology*. Vol. 5. 2, S15.

Xiao, Yandong, Marco Tulio Angulo, Jonathan Friedman, Matthew K Waldor, Scott T Weiss, and Yang-Yu Liu (2017). “Mapping the ecological networks of microbial communities.” In: *Nat. Commun.* 8.1, p. 2042.

Xu, Lizhen, Andrew D Paterson, Williams Turpin, and Wei Xu (2015). “Assessment and Selection of Competing Models for Zero-Inflated Microbiome Data.” In: *PLoS One* 10.7, e0129606.

Yang, Yuqing, Ning Chen, and Ting Chen (2017). “Inference of Environmental Factor-Microbe and Microbe-Microbe Associations from Metagenomic Data Using a Hierarchical Bayesian Statistical Model.” In: *Cell Syst* 4.1, 129–137.e5.

Yarza, Pablo, Pelin Yilmaz, Elmar Pruesse, Frank Oliver Glöckner, Wolfgang Ludwig, Karl-Heinz Schleifer, William B. Whitman, Jean Euzéby, Rudolf Amann, and Ramon Rosselló-Móra (2014). “Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences.” In: *Nature Reviews. Microbiology* 12.9, pp. 635–645.

Yilmaz, Pelin, Renzo Kottmann, Dawn Field, Rob Knight, James R Cole, Linda Amaral-Zettler, Jack A Gilbert, Ilene Karsch-Mizrachi, Anjanette Johnston, Guy Cochrane, Robert Vaughan, Christopher Hunter, Joonhong Park, Norman Morrison,

Philippe Rocca-Serra, Peter Sterk, Manimozhiyan Arumugam, Mark Bailey, Laura Baumgartner, Bruce W Birren, Martin J Blaser, Vivien Bonazzi, Tim Booth, Peer Bork, Frederic D Bushman, Pier Luigi Buttigieg, Patrick S G Chain, Emily Charlson, Elizabeth K Costello, Heather Huot-Creasy, Peter Dawyndt, Todd DeSantis, Noah Fierer, Jed A Fuhrman, Rachel E Gallery, Dirk Gevers, Richard A Gibbs, Inigo San Gil, Antonio Gonzalez, Jeffrey I Gordon, Robert Guralnick, Wolfgang Hankeln, Sarah Highlander, Philip Hugenholtz, Janet Jansson, Andrew L Kau, Scott T Kelley, Jerry Kennedy, Dan Knights, Omry Koren, Justin Kuczynski, Nikos Kyrpides, Robert Larsen, Christian L Lauber, Teresa Legg, Ruth E Ley, Catherine A Lozupone, Wolfgang Ludwig, Donna Lyons, Eamonn Maguire, Barbara A Methé, Folker Meyer, Brian Muegge, Sara Nakielny, Karen E Nelson, Diana Nemergut, Josh D Neufeld, Lindsay K Newbold, Anna E Oliver, Norman R Pace, Giriprakash Palanisamy, Jörg Peplies, Joseph Petrosino, Lita Proctor, Elmar Pruesse, Christian Quast, Jeroen Raes, Sujeevan Ratnasingham, Jacques Ravel, David A Relman, Susanna Assunta-Sansone, Patrick D Schloss, Lynn Schriml, Rohini Sinha, Michelle I Smith, Erica Sodergren, Aymé Spo, Jesse Stombaugh, James M Tiedje, Doyle V Ward, George M Weinstock, Doug Wendel, Owen White, Andrew Whiteley, Andreas Wilke, Jennifer R Wortman, Tanya Yatsunenko, and Frank Oliver Glöckner (2011). “Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications.” In: *Nat. Biotechnol.* 29.5, pp. 415–420.

Yilmaz, Pelin, Laura Wegener Parfrey, Pablo Yarza, Jan Gerken, Elmar Pruesse, Christian Quast, Timmy Schweer, Jörg Peplies, Wolfgang Ludwig, and Frank Oliver Glöckner (2013). “The SILVA and “All-species Living Tree Project (LTP)”taxonomic frameworks.” In: *Nucleic Acids Res.* 42.D1, pp. D643–D648.

Zaremba-Niedzwiedzka, Katarzyna, Eva F. Caceres, Jimmy H. Saw, Disa Bäckström, Lina Juzokaite, Emmelien Vancaester, Kiley W. Seitz, Karthik Anantharaman, Piotr Starnawski, Kasper U. Kjeldsen, Matthew B. Stott, Takuro Nunoura, Jillian F. Banfield, Andreas Schramm, Brett J. Baker, Anja Spang, and Thijs J. G. Ettema (2017). “Asgard archaea illuminate the origin of eukaryotic cellular complexity.” In: *Nature* 541.7637, pp. 353–358.

Zelezniak, Aleksej, Sergej Andrejev, Olga Ponomarova, Daniel R. Mende, Peer Bork, and Kiran Raosaheb Patil (2015). “Metabolic dependencies drive species co-occurrence in diverse microbial communities.” In: *Proceedings of the National Academy of Sciences* 112.20, pp. 6449–6454.

Zeller, Georg, Julien Tap, Anita Y Voigt, Shinichi Sunagawa, Jens Roat Kultima, Paul I Costea, Aurélien Amiot, Jürgen Böhm, Francesco Brunetti, Nina Habermann, Rajna Hercog, Moritz Koch, Alain Luciani, Daniel R Mende, Martin A Schneider, Petra Schrotz-King, Christophe Tournigand, Jeanne Tran Van Nhieu, Takuji Yamada, Jürgen Zimmermann, Vladimir Benes, Matthias Kloor, Cornelia M Ulrich, Magnus

von Knebel Doeberitz, Iradj Sobhani, and Peer Bork (2014). “Potential of fecal microbiota for early-stage detection of colorectal cancer.” In: *Molecular Systems Biology* 10.11.

Zeng, Xiang, Jean-Louis Birrien, Yves Fouquet, Georgy Cherkashov, Mohamed Jebbar, Joël Querellou, Philippe Oger, Marie-Anne Cambon-Bonavita, Xiang Xiao, and Daniel Prieur (2009). “Pyrococcus CH1, an obligate piezophilic hyperthermophile: extending the upper pressure-temperature limits for life.” In: *The ISME journal* 3.7, pp. 873–876.

Zengler, Karsten, Gerardo Toledo, Michael Rappé, James Elkins, Eric J. Mathur, Jay M. Short, and Martin Keller (2002). “Cultivating the uncultured.” In: *Proceedings of the National Academy of Sciences* 99.24, pp. 15681–15686.

Zhang, Kun, Biwei Huang, Jiji Zhang, Clark Glymour, and Bernhard Schölkopf (2017). “Causal Discovery from Nonstationary/Heterogeneous Data: Skeleton Estimation and Orientation Determination.” In: *IJCAI* 2017, pp. 1347–1353.

Zhang, Kun, Bernhard Schölkopf, Peter Spirtes, and Clark Glymour (2018a). “Learning causality and causality-related learning: some recent progress.” In: *Natl Sci Rev* 5.1, pp. 26–29.

Zhang, Mengxiang, Wei Ma, Juan Zhang, Yi He, and Juan Wang (2018b). “Analysis of gut microbiota profiles and microbe-disease associations in children with autism spectrum disorders in China.” In: *Sci. Rep.* 8.1, p. 13981.

Zhao, Yanlin, Ben Temperton, J. Cameron Thrash, Michael S. Schwalbach, Kevin L. Vergin, Zachary C. Landry, Mark Ellisman, Tom Deerinck, Matthew B. Sullivan, and Stephen J. Giovannoni (2013). “Abundant SAR11 viruses in the ocean.” In: *Nature* 494.7437, pp. 357–360.

Zhernakova, Alexandra, Alexander Kurilshikov, Marc Jan Bonder, Ettje F Tigchelaar, Melanie Schirmer, Tommi Vatanen, Zlatan Mujagic, Arnau Vich Vila, Gwen Falony, Sara Vieira-Silva, Jun Wang, Floris Imhann, Eelke Brandsma, Soesma A Jankipersadsing, Marie Joossens, Maria Carmen Cenit, Patrick Deelen, Morris A Swertz, LifeLines cohort study, Rinse K Weersma, Edith J M Feskens, Mihai G Netea, Dirk Gevers, Daisy Jonkers, Lude Franke, Yurii S Aulchenko, Curtis Huttenhower, Jeroen Raes, Marten H Hofker, Ramnik J Xavier, Cisca Wijmenga, and Jingyuan Fu (2016). “Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity.” In: *Science* 352.6285, pp. 565–569.

Zhou, Jizhong, Ye Deng, Feng Luo, Zhili He, Qichao Tu, and Xiaoyang Zhi (2010). “Functional molecular ecological networks.” In: *MBio* 1.4, e00169–10.

Zhu, Yong-Guan, Michael Gillings, Pascal Simonet, Dov Stekel, Steve Banwart, and Josep Penuelas (2017). “Microbial mass movements.” In: *Science* 357.6356, pp. 1099–1100.

Zmora, Niv, Gili Zilberman-Schapira, Jotham Suez, Uria Mor, Mally Dori-Bachash, Stavros Bashiardes, Eran Kotler, Maya Zur, Dana Regev-Lehavi, Rotem Ben-Zeev Brik, Sara Federici, Yotam Cohen, Raquel Linevsky, Daphna Rothschild, Andreas E. Moor, Shani Ben-Moshe, Alon Harmelin, Shalev Itzkovitz, Nitsan Maharshak, Oren Shibolet, Hagit Shapiro, Meirav Pevsner-Fischer, Itai Sharon, Zamir Halpern, Eran Segal, and Eran Elinav (2018). "Personalized Gut Mucosal Colonization Resistance to Empiric Probiotics Is Associated with Unique Host and Microbiome Features." In: *Cell* 174.6, 1388–1405.e21.

Zobell, Claude E and Richard Y Morita (1959). "Deep-sea bacteria." In: *Galathea Report, Copenhagen* 1, pp. 139–154.

Zoetendal, Erwin G., Antoon D. L. Akkermans, and Willem M. De Vos (1998). "Temperature Gradient Gel Electrophoresis Analysis of 16S rRNA from Human Fecal Samples Reveals Stable and Host-Specific Communities of Active Bacteria." In: *Appl. Environ. Microbiol.* 64.10, pp. 3854–3859.

Zoetendal, Erwin G., Atte von Wright, Terttu Vilpponen-Salmela, Kaouther Ben-Amor, Antoon D. L. Akkermans, and Willem M. de Vos (2002). "Mucosa-Associated Bacteria in the Human Gastrointestinal Tract Are Uniformly Distributed along the Colon and Differ from the Community Recovered from Feces." In: *Applied and Environmental Microbiology* 68.7, pp. 3401–3407.